Admin and Lecture 1: Recap of Measure Theory

David Aldous

January 16, 2018

I don't use bCourses:

- Read web page (search `Aldous 205B`)

Web page rather unorganized – some topics done by Nike in 205A – will post homeworks soon.

I am **not** following Durrett text section-by-section.

After today – math on blackboard. Start today with conceptual review of measure theory [MT] – emphasizing stuff books don't tell you.

On measure theory [MT] and probability theory [PT]

1.

> *At a purely formal level, one could call probability theory the study of measure spaces with total measure one, but that would be like calling number theory the study of strings of digits which terminate.* [Terry Tao]

My own analogy is

> *MT is like an operating system for PT*

and like any good OS it should be *transparent*. There's another Terry Tao quote about "3 stages of learning math" . . . . . .

You *shouldn't* start a research paper with $(\Omega, \mathcal{F}, \mathbb{P})$; it's just there in the background.

Historical digression: In retrospect, what was Kolmogorov's insight?

While Probability certainly involves some conceptually extra idea (relative to the rest of Mathematics), the issue (100 years ago) was whether Probability required some new technical ingredient to be added to the rest of Mathematics. Kolmogorov's achievement was the realization that it didn't. Measure theory had been recently developed to resolve the technical conflict between the intuitive idea "every region in the plane has some area" and the axioms of set theory dealing with every subset of an uncountable set. This conflict has no conceptual connection with Probability, but Kolmogorov realized that the technical machinery (involved in its resolution) of measures, measurable sets, measurable functions could be reused as an axiomatic setting for Probability.

In retrospect, because one special model within Probability is "pick a uniform random point from the unit square", it is clear that any general theory of Probability has to include measure theory, but (to reiterate) Kolmogorov's achievement was the realization that at the technical level it didn't require anything more.

2. Regarding MT:

- Not much is needed for most PT.
- Mostly it's clever definitions; only hard theorem is existence of Lebesgue measure.
- It's rather magical that integration theory works so very generally (no topology needed) . . . . . .
- . . . . . . I will explain the secret reason why.

[show picture of measurable f: forwards/backwards]

3. Two illustrative contexts where MT helps:
(a) For $\mathbb{R}$-valued X

$$\mathbb{E}X = \underbrace{\int X(\omega)P(d\omega) = \int x\mu(dx)}_{\text{(theory)}} = \underbrace{\int xf(x)dx = \sum_x x\mathbb{P}(X = x)}_{\text{(calculation).}}$$

(b) $\limsup_n X_n$ is a random variable !

(b) illustrates point that MT closed under (many) countable limits

4. You need to understand, both formally and intuitively, the relations between random variables (RV) and distributions (PM).

- Any RV (general space S) has a distribution.
- In MT one typically deals with given arbitrary PMs.
- In PT we usually think of a PM as arising as the distribution of a RV.

Think in terms of RVs rather than PMs whenever you can.

Here are some aspects of this relationship.

(a). For a PM μ on $\mathbb{R}$, take U uniform$(0,1)$ and then

$$F_\mu^{-1}(U) \sim \mu.$$

Definition: A measurable space $(S, \mathcal{S})$ is *nice* if it is isomorphic to a Borel subset B of $\mathbb{R}$.
[explain on board]

Background fact: Every space you ever encounter will be a *nice* space.

Corollary: For any PM μ on any nice space S, there is a measurable function $G_\mu : [0,1] \to S$ such that $G_\mu(U) \sim \mu$.

(But aside from $\mathbb{R}$ no canonical choice).

(b). For two PMs μ, ν on $\mathbb{R}$ the property

$$\nu(-\infty, x] \geq \mu(-\infty, x] \; \forall x$$

is equivalent to

$$\exists \; X_\mu \sim \mu, \;\; X_\nu \sim \nu \text{ such that } X_\nu \leq X_\mu$$

and a canonical choice is

$$(X_\nu, X_\mu) = (F_\nu^{-1}(U), F_\mu^{-1}(U)).$$

This is *stochastic order* $\nu \preceq \mu$.

(c). For two PMs μ, ν on $\mathbb{R}$ with finite means the property

$$\int \phi d\nu \leq \int \phi d\mu \ \forall \text{ integrable convex } \phi$$

is equivalent to

$$\exists\ X_\mu \sim \mu,\ X_\nu \sim \nu \text{ such that } \mathbb{E}(X_\mu | X_\nu) = X_\nu.$$

This is *convex order* $\nu \preceq \mu$. This result not so easy; and there is no canonical choice.

(d). For two PMs μ, ν on general space S we have *variation distance*

$$||\mu - \nu|| = \max_A |\mu(A) - \nu(A)| = \tfrac{1}{2} \sum_i |\mu_i - \nu_i| \quad (\text{ discrete })$$

with relation

$$||\mu - \nu|| = \inf\{\ \mathbb{P}(X_\mu \neq X_\nu)\ \}$$

the *inf* over joint distributions with $X_\mu \sim \mu,\ X_\nu \sim \nu$. The *inf* is attained by joint distributions with $\mathbb{P}(X_\nu = X_\mu = i) = \min(\mu_i, \nu_i)$.

Recall an application. By calculation

$$||\mathrm{Bern}(\lambda) - \mathrm{Poi}(\lambda)|| \leq \tfrac{1}{2}\lambda^2$$

which easily implies **Le Cam's theorem**:

For independent Bernouilli(p_i) RVs ξ_i,

$$||\mathrm{dist}(\sum_i \xi_i) \ - \ \mathrm{Poi}(\sum_i p_i)|| \leq \tfrac{1}{2} \sum_i p_i^2.$$

These are all examples of *coupling*, which means (in the wide sense) getting information about a relation between distributions by constructing (dependent) RVs with those distributions.

5. Alternative (better!) approach to some 205A theory.

$$\begin{array}{rcl} [0,1] & {}_{\mathbf{s}}\leftrightarrow^{\mathbf{b}} & \{0,1\}^\infty \text{ binary expansion} \\ .58231\ldots. & & .10010\ldots \end{array}$$

$$\mathbf{b}(x) = (b_1(x), b_2(x), \ldots) \text{ where } b_n(x) = [2^n x] \mod 2.$$

$$\mathbf{s}(\mathbf{b}) = \sum_n 2^{-n} b_n.$$

Here **b** takes Lebesgue measure on $[0,1]$ to fair-coin-tossing measure (product Bernoulli $(1/2)$) on $\{0,1\}^\infty$), and **s** takes it back. In RV terms, for uniform U and product Bernoulli $\mathbf{B} = (B_i)$

$$\mathbf{b}(U) =_d \mathbf{B}, \quad \mathbf{s}(\mathbf{B}) =_d U.$$

[Imagine idealized RNGs].

$$\mathbf{b}(U) =_d \mathbf{B}, \quad \mathbf{s}(\mathbf{B}) =_d U.$$

Now we do a trick. Take disjoint infinite subsets $S_1, S_2, \ldots$ of $\{1, 2, 3, \ldots\}$ and write $S_i = \{s_{i1}, s_{i2}, \ldots\}$ and define

$$U_i = \mathbf{s}(B_{s_{ij}}, j \geq 1), \; i = 1, 2, \ldots$$

This gives an infinite sequence of IID uniform$(0, 1)$ RVs.

Recall **Corollary:** For any PM μ on any nice space S, there is a measurable function $G_\mu : [0, 1] \to S$ such that $G_\mu(U) \sim \mu$.

Now we can conclude **infinite product measure** $\mu_1 \times \mu_2 \times \ldots$ exists because it is the distribution of $(G_{\mu_1}(U_1), G_{\mu_2}(U_2), \ldots)$.

This is much simpler than usual MT proofs for $\mu_1 \times \mu_2$ (though only for *nice* spaces). Note an interpretation of the Corollary: if it were false there would be PMs one could not simulate even in infinite time from an idealized RNG.

Recall

- It's rather magical that integration theory works so very generally (no topology needed) . . . . . .
- . . . . . . I will explain the secret reason why.

The secret reason is that (roughly speaking) all measure spaces $(S, \mathcal{S}, \mu)$ are the same as $([0,1], \mathcal{B}, \mathrm{Leb})$.

Another secret..

- For almost all PT you do not need to know the **proof** of existence of Lebesgue measure (via outer measures etc).

Understanding the Radon-Nikodym theorem.

Physical **density** is mass per unit volume, as a local limit for heterogeneous material. In the setting of two PMs μ, ν on $(S, \mathcal{S})$ we would like to define density as

$$f(s) = \lim_{A \downarrow \{s\}} \frac{\nu(A)}{\mu(A)}$$

but this definition doesn't work very generally. However, assuming $\mathcal{S} = \sigma(B_i, 1 \le i < \infty)$ (countably generated), we have finite fields $\mathcal{F}_n = \sigma(B_i, 1 \le i \le n)$ and we can define

$$X_n(s) = \frac{\nu(A)}{\mu(A)} \text{ where } A \text{ is atom, } s \in A$$

which is finite when $\nu \ll \mu$, that is

$$\forall A \in \mathcal{S} : \ \mu(A) = 0 \text{ implies } \nu(A) = 0.$$

Key point: (X_n) is a martingale w.r.t. $(S, \mathcal{S}, \mu)$.

$$X_n(s) = \frac{\nu(A)}{\mu(A)} \text{ where } A \text{ is atom, } s \in A$$

Key point: (X_n) is a martingale w.r.t. $(S, \mathcal{S}, \mu)$.
And assumption $\nu \ll \mu$ implies (easy: by contradiction)

$$\forall \varepsilon > 0 \; \exists \delta(\varepsilon) > 0 \text{ such that } \mu(A) < \delta(\varepsilon) \text{ implies } \nu(A) < \varepsilon$$

This implies (X_n) is uniformly integrable, and MG convergence says there is a limit function f

$$X_n(s) \to f(s) \text{ a.s., } L^1, \text{ w.r.t. } \mu$$

which has the desired "density" property

$$\nu(A) = \int_A f \; d\mu, \quad A \in \mathcal{S}.$$

So we can indeed get the density as some particular limit

$$f(s) = \lim_n \frac{\nu(A_n(s))}{\mu(A_n(s))} \; \mu - \text{ a.s.}$$

But not canonical.

Conditioning and MT

Formal definition of conditional expectation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose $X \in \mathcal{F}$ and $\mathbb{E}|X| < \infty$. Suppose $\mathcal{G}$ is a σ-field contained in $\mathcal{F}$. We say that a random variable Y is a version of $\mathbb{E}(X|\mathcal{G})$ if

(I) $Y \in \mathcal{G}$

(II) $\mathbb{E}[Y1_A] = \mathbb{E}[X1_A] \; \forall \; A \in \mathcal{G}$.

This Y exists, is integrable, and is unique (almost surely).

Then there is a "calculus of conditional expectations" which is very useful, in particular in the context of martingales, so you need to learn that. For example: if $\mathbb{E}X^2 < \infty$ then

$$\mathrm{var}\, X = \mathbb{E}\ \mathrm{var}(X|\mathcal{G}) \ + \ \mathrm{var}\ \mathbb{E}(X|\mathcal{G}).$$

[board example of use]

Much of this theory is intuitive as gambling on a fair game.
If you pay a fixed stake x to get a random return X, your gain is $X - x$, this is "fair" if $\mathbb{E}(\mathrm{gain}) = 0$, that is if $x = \mathbb{E}X$.

A sub- σ-field $\mathcal{G}$ represents "information". What is the fair stake Y if you know $\mathcal{G}$? Consider $A \in \mathcal{G}$ and the strategy:

if A occurs, stake Y, if not, don't bet.

Your gain is $(X - Y)1_A$, and to be fair we need $\mathbb{E}(\mathrm{gain}) = 0$, that is $\mathbb{E}[Y1_A] = \mathbb{E}[X1_A]$. **That's where the formal definition comes from.**

Optional sampling [stopping] theorems (OST) formalize the **conservation of fairness** principle: under technical assumptions, the overall result of any strategy involving a sequence of bets on fair games is just like a single bet on a fair game: $\mathbb{E}(\mathrm{overall\ gain}) = 0$.

Gambling is like stock market. Imagine mutual fund; can buy/sell at end of day price.

- X_n price of 1 share at end of day n
- $\mathcal{F}_n$ information at end of day n
- H_n number of shares held during day n
- $S_n =$ accrued profit at end of day n.

Related by

$$S_n - S_{n-1} = H_n(X_n - X_{n-1}).$$

From the story

- $H_n \in \mathcal{F}_{n-1}$ $\quad((H_n)$ is **predictable**)

If also (X_n) is a martingale and each H_n is bounded then

- (S_n) is a martingale.

This is the discrete analog of **stochastic calculus**. One can get many results (e.g. the **upcrossing inequality**) about (X_n) by choice of (H_n) use of OST for (S_n).