A crash course in Dirichlet processes
Part 1

Jason Swanson
UCF Probability Seminar
Jan 26, 2021

I. Background and notation

(See Foundations of Modern Probability, Kallenberg (1997) for further details)

RANDOM MEASURES

$(\Omega, \mathcal{F}, P)$: Prob. sp.

$(S, \mathcal{S})$: m'ble sp.

$M(S)$: set of σ-finite measures on $(S, \mathcal{S})$

A random measure ought to be a random variable taking values in $M(S)$. Need a σ-alg. on $M(S)$.

For $A \in \mathcal{S}$, $B \in \mathcal{R}$, Borel σ-alg. on $\mathbb{R}$

$$C_{A,B} := \{\nu \in M(S) : \nu(A) \in B\} \subset M(S)$$

cylinder set

$$\mathcal{M}(S) := \sigma(\{C_{A,B} : A \in \mathcal{S}, B \in \mathcal{R}\})$$

a σ-alg. on $M(S)$

$\mathcal{M}(S)$ is the smallest σ-alg. on $M(S)$ that makes the projections m'ble.

For $A \in \mathcal{S}$, define $\pi_A : M(S) \to \mathbb{R}$ by

projection onto A

$$\pi_A(\nu) = \nu(A)$$

$$\mathcal{M}(S) = \sigma(\{\pi_A : A \in \mathcal{S}\})$$

$$M_1(S) = \{\nu \in M(S) : \nu(S) = 1\}$$

probability measures on S

$\{1\} \in \mathcal{R}$

$$M_1(S) = C_{S,\{1\}} \qquad \therefore M_1(S) \in \mathcal{M}(S)$$

$S \in \mathcal{S}$

$$\mathcal{M}_1(S) = \{ C \cap M_1(S) : C \in \mathcal{M}(S) \}$$

a σ-alg. on $M_1(S)$;
it is $\mathcal{M}(S)$ restricted to $M_1(S)$

A random measure on S is an $M(S)$-valued random variable, i.e. a function $\mu : \Omega \to M(S)$ that is $(\mathcal{F}, \mathcal{M}(S))$-m'ble.

If $\mu(\Omega) \subset M_1(S)$, then μ is a random probability measure.

A random meas. is a special case of a "kernel".

$(T, \mathcal{T})$: m'ble space

A kernel from T to S is a m'ble function $\mu : T \to M(S)$.

Notation: given $t \in T$, $A \in \mathcal{S}$, $\mu(t, A) := (\mu(t))(A)$.

Expl: $\underset{[0,\infty)}{\underset{\cap}{t}} \overset{\mu}{\longmapsto} \underset{\mathcal{M}([0,\infty))}{\underset{\cap}{e^{-ts}\, ds}}$ (need to show this is m'ble)

Laplace transform: $F(t) = \int_{[0,\infty)} f(s)\, \mu(t, ds)$

If μ is a kernel from T to S, then μ is a **probability kernel** if $\mu(T) \subset M_1(S)$.

$$\boxed{\text{A random (probability) measure is a (probability) kernel from } \Omega \text{ to } S.}$$

<u>Checking measurability</u>:

$\mu : T \to M(S)$ is $(\mathcal{T}, \mathcal{M}(S))$-m'ble

iff $\pi_A \circ \mu$ is $(\mathcal{T}, \mathbb{R})$-m'ble $\forall A \in \mathcal{S}$

iff $t \mapsto \mu(t, A)$ is $(\mathcal{T}, \mathbb{R})$-m'ble $\forall A \in \mathcal{S}$

Alternative/equivalent formulation:

A kernel from T to S is a function

$\mu : T \times \mathcal{S} \to [0, \infty]$ such that

- $\mu(t, \cdot)$ is a σ-finite meas. $\forall t \in T$, and
- $\mu(\cdot, A)$ is m'ble $\forall A \in \mathcal{S}$.

REGULAR CONDITIONAL DISTRIBUTIONS

X: an $(S, \mathcal{S})$-valued random variable

$\mathcal{L}(X)$: the distribution (or law) of X

a prob. meas. on S

$\mu = \mathcal{L}(X) \Rightarrow$

- $P(X \in A) = \mu(A) \quad \forall A \in \mathcal{S}$
- $E[f(X)] = \int_S f(x)\,\mu(dx)$ whenever $f(X) \geq 0$ a.s. or $E|f(X)| < \infty$

Can we do the same thing with conditioning?

Is $P(X \in A \mid \mathcal{G})$ a random prob. meas?

Can we get $E[f(X) \mid \mathcal{G}]$ by integrating?

Typically, yes. But need hypotheses.

A m'ble space $(S, \mathcal{S})$ is a (standard) Borel space if $\exists$ a bijection $\varphi: S \to \mathbb{R}$ such that φ is $(\mathcal{S}, \mathcal{R})$-m'ble and φ^{-1} is $(\mathcal{R}, \mathcal{S})$-m'ble.

$$\boxed{\text{A m'ble subset of a complete, separable metric space is a standard Borel space.}}$$

key hypothesis

$(S, \mathcal{S})$: Borel sp.

X : S-valued random var.

$(T, \mathcal{T})$: m'ble sp.

Y : T-valued random var.

Theorem $\exists$ a (probability) kernel μ from T to S such that

$$P(X \in A \mid Y) = \mu(Y, A) \text{ a.s. } \forall A \in \mathcal{S}$$

the regular conditional distribution of X given Y.

Notation: $\mathcal{L}(X \mid Y) := \mu(Y)$

$\mu(Y) = \mathcal{L}(X \mid Y) \implies$

- $P(X \in A \mid Y) = \mu(Y, A)$ a.s. $\forall A \in \mathcal{S}$
- $E[f(X) \mid Y] = \int_S f(x)\, \mu(Y, dx)$ a.s.

  whenever $E|f(X)| < \infty$.

What about $X \mid \mathcal{G}$?

Surprisingly, $X \mid \mathcal{G}$ is a special case of $X \mid Y$.

$\mathcal{G}$: sub σ-alg. of $\mathcal{F}$

$(T, \mathcal{T}) := (\Omega, \mathcal{G})$

Y: identity function

$$P(X \in A \mid \mathcal{G}) = P(X \in A \mid Y)$$

So $\exists$ a prob. kernel μ from Ω to S (i.e. a random prob. meas.) such that

$$P(X \in A \mid \mathcal{G}) = \mu(A) \quad \text{a.s.} \quad \forall A \in \mathcal{S}$$

regular conditional distribution of X given $\mathcal{G}$

Notation: $\mathcal{L}(X \mid \mathcal{G}) := \mu$

$$\mu = \mathcal{L}(X \mid \mathcal{G}) \implies$$

- $P(X \in A \mid \mathcal{G}) = \mu(A)$ a.s. $\forall A \in \mathcal{S}$
- $E[f(X) \mid \mathcal{G}] = \int_S f(x)\, \mu(dx)$ a.s.

whenever $E|f(X)| < \infty$.

Helpful result:

If $Y \in \mathcal{G}$, then

$$E[f(X,Y) \mid \mathcal{G}] = \int_S f(x,Y)\,\mu(dx),$$

where $\mu = \mathcal{L}(X \mid \mathcal{G})$.

- Treat Y like a constant (since it is known), then integrate w.r.t. the (regular) conditional distribution of X given $\mathcal{G}$.
- When X & $\mathcal{G}$ are indep., $\mathcal{L}(X \mid \mathcal{G}) = \mathcal{L}(X)$ and this result is familiar to undergrads.

Final bit of notation:

If μ, ν are complex measures (e.g. finite signed measures), then

$\mu \propto \nu$ means $\exists\, c > 0$ such that

$$\mu = c\nu$$

Expl: $X_1, X_2, \ldots$ iid $\mathrm{Exp}(\lambda)$

$$d\mathcal{L}(X_j) \propto e^{-\lambda x}\, dx$$

$$d\mathcal{L}(X_1 + \cdots + X_n) \propto x^{n-1} e^{-\lambda x}\, dx$$

notationally simpler to omit normalizing constant

II. Laplace's sunrise problem

Take a pressed penny from a museum.

Flip it repeatedly.

X_n : result of n^{th} flip (0 = tails, 1 = heads)

Assumptions:

$S = \{0,1\}$, $\mathcal{S} = \mathcal{P}(S)$ (power set)

X_n is an S-valued random var.

$X_1, X_2, \ldots$ are exchangeable:

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

$\forall$ permutations σ of $\{1, \ldots, n\}$

By de Finetti's theorem (see Thm. 9.16 in Kallenberg):

$\exists$ a random prob. meas. μ on $S = \{0,1\}$

de Finetti says nothing about the distribution of μ

such that

$$\mathcal{L}(X_1, \dots, X_n \mid \mu) = \mu^n \qquad \forall n$$

More concisely:

$X := (X_1, X_2, \dots)$ (an $(S^\infty, \mathcal{S}^\infty)$-val. r.v.)

$$\mathcal{L}(X \mid \mu) = \mu^\infty$$

$\mathcal{L}(X|\mu) = \mu^\infty$ means:

- $X_1, X_2, \ldots$ are conditionally i.i.d. given μ.

- $\mathcal{L}(X_n|\mu) = \mu \quad \forall n$

  i.e. $P(X_n \in A \,|\, \mu) = \mu(A)$ a.s. $\forall n, \forall A$

μ is the (unknown) distribution of each X_n.

If we knew this distribution, flips would be i.i.d.

Without knowing μ, flips are dependent as we learn from flip to flip.

To complete the model we must choose a distribution for μ.

This is our "prior" on the unknown μ, based on whatever we may know about the pressed penny before flipping it.

μ is an $M_1(S)$-val. r.v.

$\mathcal{L}(\mu) \in M_1(M_1(S))$

We must choose an element of $M_1(M_1(S))$ to be our prior.

SIMPLIFYING μ

μ is an $M_1(S)$-val. r.v.

$$S = \{0,1\}, \quad \mathcal{S} = \mathcal{P}(S)$$

Define $\varphi : M_1(S) \longrightarrow [0,1]$ by

$$\varphi(\nu) = \nu(\{1\})$$

- φ is bijective

Borel σ-alg. on $[0,1]$

- φ is $(\mathcal{M}_1(S), \mathcal{B}_{[0,1]})$-m'ble

$\varphi = \pi_{\{1\}}$ is a projection

- φ^{-1} is $(\mathcal{B}_{[0,1]}, \mathcal{M}_1(S))$-m'ble

To check, use $\mathcal{M}_1(S) = \sigma$(cylinder sets)

Define $\theta = \varphi(\mu) = \mu(\{1\})$

θ is a $[0,1]$-val. r.v.

$\sigma(\theta) = \sigma(\mu)$

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \mu)$$

$$= \mu(x_1)\mu(x_2)\cdots\mu(x_n)$$

$$= \theta^a (1-\theta)^b$$

θ is the unknown probability of heads.

$a = |\{j : x_j = 1\}|$

$b = |\{j : x_j = 0\}|$

We must choose a prior distribution for θ.

$$\mathcal{L}(\mu) \in M_1(M_1(\{0,1\}))$$

$$\mathcal{L}(\theta) \in M_1([0,1])$$

We must choose a prob. meas. on $[0,1]$ as our prior on θ.

Any will do, but suppose we choose one with a density.

$$d\mathcal{L}(\theta) = f(t)\,dt$$

posterior distribution of θ given first n observations

$$\mathcal{L}(\theta \mid X_1, \ldots, X_n) = ?$$

Basic calculations give

$$d\mathcal{L}(\theta \mid X_1, \dots, X_n) \propto \underbrace{t^N (1-t)^M}_{\text{beta distribution}} f(t)\, dt,$$

$$N = |\{j : X_j = 1\}| = \sum_{j=1}^n X_j$$

$$M = |\{j : X_j = 0\}| = n - N$$

The beta distribution is a special case of the Dirichlet distribution.

The Dirichlet distribution is a discrete version of the Dirichlet process.

BETA DISTRIBUTION

The beta distribution with parameters $\alpha > 0$ and $\beta > 0$, denoted $\mathrm{Beta}(\alpha, \beta)$, is the probability measure on $[0,1]$ proportional to

$$t^{\alpha-1}(1-t)^{\beta-1}\,dt.$$

$\alpha = 1$, $\beta = 1$ is the uniform distribution

If $\mu = \mathrm{Beta}(\alpha, \beta)$, then

$$d\mu = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

$$= \frac{1}{B(\alpha,\beta)}\, t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

beta function

mean of $\mathrm{Beta}(\alpha,\beta)$ is $\dfrac{\alpha}{\alpha+\beta}$

$$d\mathcal{L}(\theta) = f(t)\,dt \quad \Longrightarrow$$

$$d\mathcal{L}(\theta \mid X_1, \dots, X_n) \propto t^N (1-t)^M f(t)\,dt,$$

$$N = \sum_{j=1}^n X_j \,, \quad M = n - N$$

If the prior is beta, then the posterior is beta.

$$\mathcal{L}(\theta) = \mathrm{Beta}(\alpha, \beta) \Longrightarrow$$

$$d\mathcal{L}(\theta \mid X_1, \dots, X_n) \propto t^N (1-t)^M t^{\alpha-1} (1-t)^{\beta-1} dt$$

$$= t^{\alpha+N-1} (1-t)^{\beta+M-1} dt$$

$$\Rightarrow \mathcal{L}(\theta \mid X_1, \dots, X_n) = \mathrm{Beta}(\alpha+N, \beta+M)$$

Expl Flip the pressed penny n times. Suppose we get s heads. What is the probability of heads on the $(n+1)^{th}$ flip? Take $\mathcal{L}(\theta)$ to be uniform?

$$P(X_{n+1}=1 \mid N=s) = ?$$

$$P(X_{n+1}=1 \mid N) = E[P(X_{n+1}=1 \mid \theta, N) \mid N]$$

$$= E[P(X_{n+1}=1 \mid \theta) \mid N]$$

$$= E[\theta \mid N]$$

$$= E[E[\theta \mid X_1, \dots, X_n] \mid N]$$

$$\mathcal{L}(\theta) = \text{Beta}(1,1) \implies$$

$$\mathcal{L}(\theta \mid X_1, \ldots, X_n) = \text{Beta}(1+N, 1+M)$$

$$\text{mean of Beta}(\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$$

$$\therefore E[\theta \mid X_1, \ldots, X_n] = \frac{N+1}{n+2}$$

$N + M = n$

$$\therefore P(X_{n+1} = 1 \mid N) = E\big[E[\theta \mid X_1, \ldots, X_n] \mid N\big]$$

$$= \frac{N+1}{n+2}$$

$$P(X_{n+1} = 1 \mid N = s) = \frac{s+1}{n+2}. \quad \square$$