Conditional Expectation

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

10 Conditioning

Frequently in probability and (especially Bayesian) statistics we wish to find the probability of some event A or the expectation of some random variable X, *conditionally* on some body of information— such as the occurrence of another event B or the value of another random variable Z (or collection of them $\{Z_\alpha\}$). In elementary probability we encounter the usual formulas for conditional probabilities and expectations

$$\mathsf{P}[A \mid B] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]} \qquad \mathsf{E}[X \mid Z] = \begin{cases} \frac{\int x\, f(x,Z)\, dx}{\int f(x,Z)\, dx} & X, Z \text{ jointly continuous} \\ \frac{\sum x\, f(x,Z)}{\sum f(x,Z)} & X, Z \text{ discrete} \end{cases}$$

but this notion breaks down either for distributions which are *not* jointly absolutely continuous or discrete, and also when we wish to condition on the value of infinitely-many (even uncountably-many) random variables $\{Z_\alpha\}$, as we will when we consider stochastic processes. There simply is no such thing as a joint density function for an infinite collection $\{Z_\alpha\}$, even if each finite set has an absolutely continuous joint distribution.

Since "information" in probability theory is represented by σ-algebras (here $\sigma\{B\}$ or $\sigma\{Z_\alpha\}$), what we need are ways to express, interpret, and compute *conditional* probabilities of events and *conditional* expectations of random variables, given σ-algebras. As a bonus, this will unify the notions of conditional probability and conditional expectation, for distributions that are discrete or continuous or neither. First, a tool to help us.

10.1 Lebesgue's Decomposition

Let μ and λ be two positive σ-finite measures on the same measurable space $(\Omega, \mathcal{F})$. Call μ and λ *equivalent*, and write $\mu \equiv \lambda$, if they have the same null sets— so the notion of "a.e." is the same for both. More generally, we call λ *absolutely continuous* (AC) w.r.t. μ, and write $\lambda \ll \mu$, if $\mu(A) = 0$ implies $\lambda(A) = 0$, *i.e.*, if every μ-null set is also λ-null (so $\lambda \equiv \mu$ if and only if $\lambda \ll \mu$ and $\mu \ll \lambda$). We call μ and λ *mutually singular*, and write $\mu \perp \lambda$, if for some

disjoint sets $A, B \in \mathcal{F}$ we have $\mu(A^c) = 0$ and $\lambda(B^c) = 0$, so μ and λ are "concentrated" on disjoint sets.

For example— if $\lambda(A) = \int_A f(x)\mu(dx)$ for some $\mathcal{F}$-measurable function $f : \Omega \to \mathbb{R}_+$, then $\lambda \ll \mu$; if $f > 0$ μ-*a.s*, then also $\mu(A) = \int_A f(x)^{-1}\lambda(dx)$ and $\mu \equiv \lambda$. If for some other σ-finite measure ν and some $\mathcal{F}$-measurable $f, g : \Omega \to \mathbb{R}_+$ we set

$$\mu(A) := \int_A f(x)\nu(dx) \qquad \lambda(A) := \int_A g(x)\nu(dx)$$

then $\mu \perp \lambda$ if $f(x)g(x) = 0$ for ν-a.e. $x \in \Omega$. The functions f and g are called the densities of μ and λ *with respect to* ν, generalizing the familiar idea of density functions w.r.t. Lebesgue measure.

Theorem 1 (Lebesgue Decomposition) *Let μ, λ be two σ-finite measures on a countably-generated*[1] *measurable space $(\Omega, \mathcal{G})$. Then there exist a unique pair λ_a, λ_s of σ-finite measures on $(\Omega, \mathcal{G})$ and a unique $\mathcal{G}$-measurable function $Y \geq 0$ such that:*

$$\begin{aligned} \lambda &= \lambda_a + \lambda_s \\ \lambda_a \ll \mu, &\qquad \lambda_s \perp \mu \\ \lambda_a(G) = \int_G Y(\omega)\mu(d\omega), &\qquad G \in \mathcal{G}. \end{aligned}$$

Proof Sketch. First take λ finite. Set

$$\mathcal{H} := \left\{ h \in L_1(\Omega, \mathcal{G}, \mu) : \ h \geq 0, \ (\forall G \in \mathcal{G}) \ \int_G h \, d\mu \leq \lambda(G) \right\}$$

Show that $\mathcal{H}$ is closed under maxima, then find simple $\{h_n \geq 0\} \subset L_1$ such that

$$\sup \left\{ \int h_n \, d\mu : \ n \in \mathbb{N} \right\} = \sup \left\{ \int h \, d\mu : \ h \in \mathcal{H} \right\}$$

and set $h := \sup h_n$ and $Y := h\mathbf{1}_{\{h<\infty\}}$. Now verify the statement of the Theorem. The extension to σ-finite λ is straightforward (why?).

To show uniqueness, suppose $\lambda = \lambda_a + \lambda_s = \lambda'_a + \lambda'_s$ are two decompositions with $\lambda_a(d\omega) = Y(\omega)\mu(d\omega)$ and $\lambda'_a(d\omega) = Y'(\omega)\mu(d\omega)$. Find a single disjoint pair $A, B \in \mathcal{G}$ with $\lambda_s(A^c) = \lambda'_s(A^c) = 0$ and $\mu(B^c) = 0$, and set $G := \{\omega : \ (Y - Y') > 0\}$. Show that if $\mu(G) = \mu(G \cap B) > 0$ then also $\lambda_a(G) > \lambda'_a(G)$, but $\lambda_s(B) = \lambda'_s(B) = 0$, a contradiction. ∎

If $\mu(dx) = dx$ is Lebesgue measure on $\mathbb{R}^d$, for example, then this decomposes any probability distribution λ into an absolutely continuous part $\lambda_a(dx) = Y(x)\,dx$ with pdf Y and a singular part $\lambda_s(dx)$ (the sum of the singular-continuous and discrete components).

[1] For example, the Borel sets on any complete separable metric space.

When $\lambda \ll \mu$ (so $\lambda_a = \lambda$ and $\lambda_s = 0$) the Radon-Nikodym derivative is often denoted

$$Y = \frac{d\lambda}{d\mu} = \frac{\lambda(d\omega)}{\mu(d\omega)},$$

and extends the idea of "density" from densities with respect to Lebesgue measure to those with respect to an arbitrary "reference" (or "base" or "dominating") measure μ. For example, the pmf $f(x) = \mathsf{P}[X = x]$ of an integer-valued random variable X may now be viewed as the pdf of its distribution with respect to counting measure on $\mathbb{Z}$, so families of discrete distributions now have pdf's (if they take values in a common countable set), and random variables with mixed distributions (truncated normals, for example) have density functions with respect to a dominating measure that includes point masses where the distributions have atoms, and Lebesgue measure where they are absolutely continuous. With respect to the finite base measure $\lambda(A) := \sum\{1/k! : \ k \in A\}$ on the nonnegative integers $\mathbb{N}_0$, for example, the $\mathsf{Po}(\lambda)$ distribution has pdf $f(k) = \lambda^k e^{-\lambda}$.

To explore further conditioning we apply Lebesgue's decomposition in a quite different way, with $\mu = \mathsf{P}$ a probability measure on $(\Omega, \mathcal{F}, \mathsf{P})$ and $\lambda(d\omega) = X\, d\mathsf{P}$ for some $X \in L_1$ a σ-finite measure to prove the important:

10.2 The Radon-Nikodym Theorem

Theorem 2 (Radon-Nikodym) *Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ an integrable random variable, and $\mathcal{G} \subset \mathcal{F}$ a sub-σ-algebra. Then there exists a unique integrable $Y \in L_1(\Omega, \mathcal{G}, \mathsf{P})$, which we will denote $Y = \mathsf{E}[X \mid \mathcal{G}]$ and call a "conditional expectation of X, given $\mathcal{G}$," that satisfies for every $G \in \mathcal{G}$:*

$$(\forall G \in \mathcal{G}) \qquad \mathsf{E}\ (X - Y)\mathbf{1}_G = 0$$

The important feature to notice is that Y must be $\mathcal{G}$-measurable, which may be hard to achieve if $\mathcal{G}$ is much smaller than $\mathcal{F}$. In some sense Y is the best possible $\mathcal{G}$-measurable approximation to X.

Proof. First take X to be non-negative, $X \geq 0$. The measure $\mathcal{P}$, initially defined on all of $\mathcal{F}$, can also be viewed as a probability measure on the smaller σ-algebra $\mathcal{G} \subset \mathcal{F}$. Define another measure λ on $\mathcal{G}$ (*not* on all of $\mathcal{F}$) by

$$\lambda(G) := \mathsf{E}X\,\mathbf{1}_G = \int_G X(\omega)\,\mathsf{P}(d\omega), \qquad G \in \mathcal{G}.$$

This is bounded (since $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$) and positive (since $X \geq 0$), so by Theorem 1 (applied on $(\Omega, \mathcal{G}, \mathsf{P})$, *not* $(\Omega, \mathcal{F}, \mathsf{P})$) we can write $\lambda = \lambda_a + \lambda_s$ with $\lambda_a \ll \mathsf{P}$, $\lambda_s \perp \mathsf{P}$, and $\lambda_a(G) = \int_G Y\, d\mathsf{P}$ for some $Y \in L_1(\Omega, \mathcal{G}, \mathsf{P})$. But $\lambda \ll \mathsf{P}$ by construction, so (by uniqueness) $\lambda_s = 0$, $\lambda_a = \lambda$, and the Theorem follows.

For general X, consider separately the positive and negative parts $X_+ := \max(X, 0)$ and $X_- := \max(-X, 0)$ and set $Y := Y_+ - Y_-$. ∎

For events $A \in \mathcal{F}$ and sub-σ-algebras $\mathcal{G} \subseteq \mathcal{F}$ we denote the *conditional probability of A, given* $\mathcal{G}$ by

$$\mathsf{P}[A \mid \mathcal{G}] = \mathsf{E}\left[\mathbf{1}_A \mid \mathcal{G}\right],$$

a $\mathcal{G}$-measurable random variable (*not* a numerical constant) taking values in $[0, 1]$.

Of course X itself has the property that its integrals over events $G \in \mathcal{G}$ coincide with those of X— the point is that $Y = \mathsf{E}[X \mid \mathcal{G}]$ is a $\mathcal{G}$*-measurable approximation* to X (*i.e.*, one that depends only on the "information" encoded in $\mathcal{G}$) with this property. As we'll see below, if $\mathcal{F} \subseteq \mathcal{G}$ (or, more generally, if X is $\mathcal{G}$-measurable, so $\sigma(X) \subseteq \mathcal{G}$) then the best $\mathcal{G}$-measurable approximation is $\mathsf{E}[X \mid \mathcal{G}] = X$ itself. At the other extreme, if X is independent of $\mathcal{G}$, then one can do no better than the constant random variable $\mathsf{E}[X \mid \mathcal{G}] \equiv \mathsf{E}X$.

10.2.1 Key Example: Countable Partitions

If $\mathcal{G} = \sigma\{\Lambda_n\}$ for a finite or countable partition $\{\Lambda_n\} \subset \mathcal{F}$ (so $\Lambda_m \cap \Lambda_n = \emptyset$ for $m \neq n$ and $\Omega = \cup\Lambda_n$), then for any $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$,

$$\mathsf{E}[X \mid \mathcal{G}] = \sum \mathbf{1}_{\Lambda_n} \mathsf{E}_{\Lambda_n}[X] = \sum \mathbf{1}_{\Lambda_n}(\omega) \frac{1}{\mathsf{P}[\Lambda_n]} \mathsf{E}[X\, \mathbf{1}_{\Lambda_n}]$$

is constant on partition elements and equal there to the P-weighted average value of X (omit from the sum any term with $\mathsf{P}[\Lambda_n] = 0$).

In particular— let $(\Omega, \mathcal{F}, \mathsf{P})$ be the unit interval with Lebesgue measure, and let $\mathcal{G}_n := \sigma\{(i/2^n, j/2^n]\}$, $0 \le i < j \le 2^n$. Note that $\mathcal{G}_n \subset \mathcal{G}_m$ for $n \le m$ and that $\mathcal{F} = \bigvee \mathcal{G}_n$. Then for any $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$,

$$X_n := \mathsf{E}[X \mid \mathcal{G}_n] = 2^n \int_{i/2^n}^{(i+1)/2^n} X(v)\, dv, \qquad i/2^n < \omega \le (i+1)/2^n,\ 0 \le i < 2^n.$$

For $m > n$ the conditional expectation $Y := \mathsf{E}[X_m \mid \mathcal{G}_n]$ would have to be $\mathcal{G}_n$-measurable and satisfy $0 = \mathsf{E}(X_m - Y)\mathbf{1}_G$ for each $G \in \mathcal{G}_n$— but X_n satisfies both those criteria, since $\mathsf{E}(X_n - X)\mathbf{1}_G = 0 = \mathsf{E}(X_m - X)\mathbf{1}_G$ for all $G \in \mathcal{G}_n \subset \mathcal{G}_m$. This is our first example of a *martingale*, a sequence of random variables $X_n \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ with the property that $X_n = \mathsf{E}[X_m \mid \mathcal{G}_n]$ for $n \le m$; we'll see more soon. What happens to X_n as $n \to \infty$?

10.2.2 Properties:

- The conditional expectation is *almost* unique: if Y_1 and Y_2 are each $\mathcal{G}$-measurable and for some $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and all $G \in \mathcal{G}$ satisfy

$$\mathsf{E}(X - Y_1)\mathbf{1}_G = 0 = \mathsf{E}(X - Y_2)\mathbf{1}_G,$$

then each may be called "$\mathsf{E}[X \mid \mathcal{G}]$" but they may not be equal for all $\omega \in \Omega$. The difference $(Y_1 - Y_2)$ is $\mathcal{G}$ measurable and is zero almost-surely, since $\mathsf{E}|Y_1 - Y_2| = \int_G (Y_1 - Y_2) d\mathsf{P} + \int_{G^c} (Y_2 - Y_1) d\mathsf{P} = 0$ for $G := \{Y_1 > Y_2\} \in \mathcal{G}$, but still may not vanish for all $\omega \in \Omega$. Thus one speaks of "a" conditional expectation rather than "the" conditional expectation.

- If $X = \mathbf{1}_A$ and if $\mathcal{G} = \sigma\{B\}$ for some $A, B \in \mathcal{F}$ with $0 < \mathsf{P}[B] < 1$,

$$\mathsf{P}[A \mid \mathcal{G}](\omega) = \mathsf{E}[\mathbf{1}_A \mid \sigma(B)](\omega) = \begin{cases} \mathsf{P}[A \cap B] \,/\mathsf{P}[B] & \omega \in B \\ \mathsf{P}[A \cap B^c]/\mathsf{P}[B^c] & \omega \notin B \end{cases}$$

  Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the notion of the conditional probability of one event A given another B (or its complement B^c).

- More generally, If $X \in L_1$ and if $\mathcal{G} = \sigma\{G_i\}$ for some (finite or countable) measurable partition $\{G_i\} \subset \mathcal{F}$, then

$$\mathsf{E}[X \mid \mathcal{G}](\omega) = \sum \mathbf{1}_{G_i}(\omega) \frac{1}{\mathsf{P}(G_i)} \int_{G_i} X(\omega') \, \mathsf{P}(d\omega')$$

  is the weighted average of X over the partition element that contains ω.

- If X is a RV on $(\Omega, \mathcal{F}, \mathsf{P})$ and $\mathcal{G} \subset \mathcal{F}$, the function-valued random variable

$$F_X(x \mid \mathcal{G}) := \mathsf{P}[X \le x \mid \mathcal{G}]$$

  (a $(\mathcal{B} \times \mathcal{G})$-measurable function of $x \in \mathbb{R}$ and $\omega \in \Omega$) is a *conditional CDF of X, given $\mathcal{G}$*. It satisfies (almost surely) the usual CDF properties— non-decreasing, right continuous, with limits 0 and 1 as $x \to -\infty$ and $x \to \infty$, respectively. If this is absolutely continuous, *i.e.*, if there is a random Borel function (*i.e.*, $(\mathcal{B} \times \mathcal{G})$-measurable) $f_X(\xi \mid \mathcal{G}) \ge 0$ on $\mathbb{R} \times \Omega$ such that $(\forall x \in \mathbb{R})$

$$F_X(x \mid \mathcal{G}) = \int_{-\infty}^{x} f_X(\xi \mid \mathcal{G}) \, d\xi \quad a.s.,$$

  then $f_X(\xi \mid \mathcal{G})$ is called a *conditional pdf of X given $\mathcal{G}$* and for any Borel function g such that $g(X) \in L_1$ we can evaluate conditional expectations by:

$$\mathsf{E}\big[g(X) \mid \mathcal{G}\big] = \int_{\mathbb{R}} g(x) f_X(x \mid \mathcal{G}) \, dx.$$

- Notation: If $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and $\mathcal{G} = \sigma(Z)$ for some RV Z, then "$\mathsf{E}[X \mid Z]$" is a short way of writing $\mathsf{E}[X \mid \sigma(Z)]$ or $\mathsf{E}[X \mid \mathcal{G}]$. Recall Y is $\sigma(Z)$-measurable if and only if it can be written in the form $Y = \phi(Z)$ for some Borel function $\phi : \mathbb{R} \to \mathbb{R}$ (obvious for simple RVs Y, then take monotone limits).

- In particular, if $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and Z is any RV, then for $\mathcal{G} = \sigma(Z)$ the quantity

$$F_X(x \mid Z) := \mathsf{P}[X \le x \mid Z]$$

is a Borel-measurable function of x and Z, called the *conditional CDF of X given Z*. If that function is absolutely continuous in the x variable almost-surely, *i.e.*, if for some Borel function $f_X(\cdot \mid \cdot) \ge 0$ on $\mathbb{R}^2$

$$F_X(x \mid Z) = \int_{-\infty}^{x} f_X(\xi \mid Z)\, d\xi \qquad a.s.,$$

then $f_X(x \mid Z)$ is the *conditional pdf of X, given Z* and for any Borel g s.t. $g(X) \in L_1$,

$$\mathsf{E}\big[g(X) \mid Z\big] = \int_{\mathbb{R}} g(x) f_X(x \mid Z)\, dx.$$

- If $X, Z \sim f(x, z)$ are jointly absolutely-continuous, $g(X) \in L_1$, and $\mathcal{G} = \sigma(Z)$,

$$\mathsf{E}[g(X) \mid Z] := \mathsf{E}[g(X) \mid \sigma(Z)] = \int g(x) \left\{ \frac{f(x, Z)}{\int f(\xi, Z)\, d\xi} \right\} dx.$$

Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the elementary notion of conditional expectation (given an RV Z), with conditional pdf given explicitly by

$$f_X(x \mid Z) = \frac{f(x, Z)}{\int f(\xi, Z)\, d\xi}.$$

What if X and Z are both discrete? What if just one is discrete? What if Z is a vector?

To prove this property, first recall that a random variable is $\mathcal{G} = \sigma(Z)$ measurable if and only if it is a Borel function of Z. Apply this to write $\mathsf{E}[g(X) \mid \mathcal{G}] = \phi(Z)$; then for $G \in \mathcal{G}$, solve the equation $0 = \mathsf{E}\mathbf{1}_G[\phi(Z) - g(X)]$ for $\phi(Z)$.

- If $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and if $X \perp\!\!\!\perp \mathcal{G}$ then

$$\mathsf{E}[X \mid \mathcal{G}] \equiv \mathsf{E}X.$$

In particular, $\mathsf{E}[X \mid \{\Omega, \emptyset\}] = \mathsf{E}X$. Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the elementary notion of expectation.

- If $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and if $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then

$$\mathsf{E}[X \mid \mathcal{H}] = \mathsf{E}\big[\, \mathsf{E}[X \mid \mathcal{G}] \,\big|\, \mathcal{H}\big]$$

This is called the "tower" (or sometimes "smoothing" or "telescoping") property of conditional expectation. It's especially useful when we have entire nested families (called *filtrations*) of σ-algebras $\{\mathcal{F}_n\}$ with $n \le m \Rightarrow \mathcal{F}_n \subseteq \mathcal{F}_m$; for example, $\mathcal{F}_n := \sigma\{X_j : \ j \le n\}$ for a family $\{X_n\}$ of (non-necessarily-independent) random variables.

- A common use of the tower property is the calculation for $\mathcal{G}$-measurable X with $Y,\ XY \in L_1(\Omega, \mathcal{F}, \mathsf{P})$, that

$$\mathsf{E}[XY \mid \mathcal{G}] = X\ \mathsf{E}[Y \mid \mathcal{G}].$$

  Thus $\mathcal{G}$-measurable RVs like X can be pulled out of conditional expectations just like constants. The case $\mathcal{G} = \sigma(X)$ is most common: $\mathsf{E}[XY \mid X] = X\ \mathsf{E}[Y \mid X]$.

- If X and $\{Y_n\}$ are jointly Gaussian, then $\mathsf{E}[X \mid \sigma\{Y_n\}]$ is the orthogonal projection of X onto the linear span of $\{Y_n\}$ in the Hilbert space $L_2(\Omega, \mathcal{F}, \mathsf{P})$. This is usually the easiest way to compute conditional expectations in multivariate normal examples. Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the notion of orthogonal projection, for Gaussian RVs. This does not generalize to non-Gaussian L_2 variables; however,

- L_2 prediction: For $X \in L_2$, $\mathsf{E}[X \mid \mathcal{G}]$ minimizes $\|X - Y\|_2^2$ among all $\mathcal{G}$-measurable Y.

- Martingales: Let $\{X_n\} \subset L_1(\Omega, \mathcal{F}, \mathsf{P})$ be iid with means $\mu = \mathsf{E}[X_n]$ and set $S_n := \sum_{j \le n} X_j$ and $\mathcal{G}_n := \sigma\{X_1, ..., X_n\}$. Then for $n \le m$,

$$\mathsf{E}[S_m \mid \mathcal{G}_n] = S_n + (m - n)\mu;$$

  it follows that $(S_n - n\mu)$ is a *martingale*. If $\{X_n\} \subset L_2(\Omega, \mathcal{F}, \mathsf{P})$, set $\sigma^2 := \mathsf{V}X_n$ and check that $(S_n - n\mu)^2 - n\sigma^2$ is also a martingale.

- Monotonicity: If $X \ge Z$ *a.s.*, then $\mathsf{E}[X \mid \mathcal{G}] \ge \mathsf{E}[Z \mid \mathcal{G}]$ *a.s.* for any $\mathcal{G} \subset \mathcal{F}$. To see this, set $Y := \mathsf{E}[X - Z \mid \mathcal{G}]$ and $G := \{\omega :\ Y < 0\}$. Since $G \in \mathcal{G}$ and $(X - Z) \ge 0$ *a.s.*, $\mathsf{E}[Y\mathbf{1}_G] = \mathsf{E}(X - Z)\mathbf{1}_G \ge 0$ so $\mathsf{P}[Y < 0] = \mathsf{P}\big[\mathsf{E}[X \mid \mathcal{G}] < \mathsf{E}[Z \mid \mathcal{G}]\big] = 0$ as claimed.

- Conditional Mean/Variance Formula: If $X \in L_2(\Omega, \mathcal{F}, \mathsf{P})$ and $Y := \mathsf{E}[X \mid \mathcal{G}]$,

$$\mathsf{V}[X] = \mathsf{E}\Big\{\mathsf{E}\big[(X - Y)^2 \mid \mathcal{G}\big]\Big\} + \mathsf{V}[Y].$$

  Thus the variance of X is the mean of the conditional variance plus the variance of the conditional mean. This elegant formula is worth remembering.

- All the usual integration tools and inequalities— DCT, MCT, Fatou, Jensen, Hölder, Minkowski, Markov, Chebychev, *etc.*— have *conditional* versions as well. For example, for $X \in L_1$ and convex $\phi(\cdot)$ with $\phi(X) \in L_1$,

$$\phi\big(\mathsf{E}[X \mid \mathcal{G}]\big) \le \mathsf{E}[\phi(X) \mid \mathcal{G}]\ \ a.s$$

  Note both sides are $\mathcal{G}$-measurable *random variables* now, not constants as in the familiar Jensen inequality, so the "almost surely" qualification is needed.
  If $0 \le X_n \uparrow X$ in probability, for another example, then

$$\mathsf{E}[X_n \mid \mathcal{G}] \to \mathsf{E}[X \mid \mathcal{G}]\ \ a.s.,$$

and also $\mathsf{E}\big[|X_n - X| \,\big|\, \mathcal{G}\big] \to 0$ *a.s*, a conditional generalization of Lebesgue's MCT. This MCT can be used to prove a conditional Fatou's Lemma for $X_n \ge 0$:

$$\mathsf{E}\big[\liminf_{n\to\infty} X_n \mid \mathcal{G}\big] \le \liminf_{n\to\infty} \mathsf{E}[X_n \mid \mathcal{G}] \; a.s.$$

If $X_n \to X (pr.)$ and $|X_n| \le Y \in L_1$ *a.s.*, then again

$$\mathsf{E}[X_n \mid \mathcal{G}] \to \mathsf{E}[X \mid \mathcal{G}] \quad \text{and} \quad \mathsf{E}\big[|X_n - X| \mid \mathcal{G}\big] \to 0 \qquad a.s.,$$

a conditional version of Lebesgue's DCT. To prove this, just apply the conditional Fatou's lemma to the nonnegative RVs $Y + X_n$ and $Y - X_n$ to see

$$\begin{aligned}
\mathsf{E}[Y + X \mid \mathcal{G}] &= \mathsf{E}[\liminf_{n\to\infty}(Y + X_n) \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathsf{E}[Y + X_n \mid \mathcal{G}] \text{ and} \\
\mathsf{E}[Y - X \mid \mathcal{G}] &= \mathsf{E}[\liminf_{n\to\infty}(Y - X_n) \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathsf{E}[Y - X_n \mid \mathcal{G}].
\end{aligned}$$

Upon subtracting $\mathsf{E}[Y \mid \mathcal{G}]$ and changing signs for the second equation, we conclude

$$\mathsf{E}[X \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathsf{E}[X_n \mid \mathcal{G}] \le \limsup_{n\to\infty} \mathsf{E}[X_n \mid \mathcal{G}] \le \mathsf{E}[X \mid \mathcal{G}].$$

10.3 An Example: Poisson Processes

Imagine that we are fishing at a site where, on average, we catch λ fish per hour. Imagine further that the numbers of fish we catch in disjoint time intervals are independent. As we shall see below, it follows that

- The *number* of fish caught by time $t > 0$ has a Poisson distribution $N_t \sim \mathsf{Po}(\lambda t)$ with mean λt and, moreover, for any sequence of times $0 \le t_1 \le t_2 \le t_3 \le \cdots \le t_n$, the numbers of fish caught in intervals $(t_j, t_{j+1}]$ are independent Poisson RVs:
$$\big(N_{t_{j+1}} - N_{t_j}\big) \overset{\text{ind}}{\sim} \mathsf{Po}\big(\lambda(t_{j+1} - t_j)\big)$$
- The *times* of catching the first, second, third, *etc.* fish $T_1, T_2, T_3, \cdots$ have iid increments
$$(T_{j+1} - T_j) \overset{\text{iid}}{\sim} \mathsf{Ex}(\lambda).$$
and Gamma marginals $T_j \sim \mathsf{Ga}(j, \lambda)$.

Proof. Fix $t > 0$. For large n, the number of fish caught in a short interval of time $(\frac{j}{n}t, \frac{j+1}{n}t]$ will be zero or one with high probability (it's hard to catch two fish in a millisecond!), so we can view the number N_t caught in time t as (approximately) the total number of successes in a fixed number n of independent trials, all with the same probability of success— so N_t has, approximately, the $\mathsf{Bi}(n, p_n)$ distribution for some success probability $0 < p_n < 1$. But the expected number caught will be $\mathsf{E}N_t = \lambda t = n \times p_n$, so $p_n = \lambda t/n$ and for any $k \in \mathbb{Z}_0$

$$\begin{aligned}
\mathsf{P}[N_t = k] &\approx \binom{n}{k}\Big(\frac{\lambda t}{n}\Big)^k\Big(1 - \frac{\lambda t}{n}\Big)^{n-k} \\
&= \frac{\overbrace{n(n-1)\cdots(n-k+1)}^{k \text{ terms}}}{k! \quad \underbrace{(n)(n)\cdots(n)}_{k \text{ terms}}}(\lambda t)^k\Big(1 - \frac{\lambda t}{n}\Big)^n\Big(1 - \frac{\lambda t}{n}\Big)^{-k} \\
&\to \frac{(\lambda t)^k}{k!}\, e^{-\lambda t} \quad \text{as } n \to \infty, \text{ so} \quad N_t \sim \mathsf{Po}(\lambda t).
\end{aligned}$$

∎

The event $[T_1 > t]$ that it takes longer than t hours to catch the first fish is the same as the event $[N_t = 0]$ that no fish are caught by time t, so

$$\mathsf{P}[T_1 > t] = \mathsf{P}[N_t = 0] = \frac{(\lambda t)^0}{0!}e^{-\lambda t} = e^{-\lambda t}$$

for $t > 0$ and $T_1 \sim \mathsf{Ex}(\lambda)$. Similarly each increment $(T_{j+1} - T_j) \sim \mathsf{Ex}(\lambda)$, and all are independent, so (marginally) the event times have Gamma distributions $T_j \sim \mathsf{Ga}(j, \lambda)$. BTW, this gives a simple way to compute Gamma probabilities. For $t > 0$, $\lambda > 0$, and $j \in \mathbb{N}$,

$$\texttt{pgamma}(\texttt{t}, \texttt{j}, \texttt{lambda}, \texttt{lower} = \texttt{F}) = \mathsf{P}[T_j > t] = \mathsf{P}[N_t < j] = e^{-\lambda t}\sum_{k<j}(\lambda t)^k/k!.$$

With this example in hand, we can compute lots of conditional expectations. For example, set $\mathcal{F}_t := \sigma\{N_s : \ s \le t\}$ and find:

- For $t > 0$, what is $\mathsf{P}[T_1 \le t \mid T_2]$? What is $\mathsf{P}[T_2 > t \mid T_1]$?
- For $i < j$, what is $\mathsf{E}[T_j \mid T_i]$? What is $\mathsf{E}[T_i \mid T_j]$?
- For $s < t$, what is $\mathsf{E}[N_t \mid N_s]$? What is $\mathsf{E}[N_t \mid \mathcal{F}_s]$?
- For $s < t$, what is $\mathsf{E}[\exp(i\omega N_t) \mid \mathcal{F}_s]$?
- For $s < t$, what is $\mathsf{E}[N_s \mid N_t]$? What is $\mathsf{E}[N_s \mid \mathcal{F}_t]$?
- For $s < t$, what is $\mathsf{E}[{N_t}^2 \mid \mathcal{F}_s]$?

How to compute conditional expectations

One way to solve in general for conditional expectations $Y := \mathsf{E}[g(X) \mid \mathcal{G}]$ is to begin by expressing what all possible $\mathcal{G}$-measurable random variables look like— for example, if $\mathcal{G} = \sigma(Z)$ then $Y = \phi(Z)$ for some Borel ϕ— and then find elements $G \in \mathcal{G}$ for which it is easy to evaluate $\mathsf{E}\big[\big(g(X) - Y\big)\mathbf{1}_G\big]$ and see what are the conditions on Y (*e.g.*, on ϕ if $\mathcal{G} = \sigma(Z)$ so $Y = \phi(Z)$) that make it vanish. A second approach is to identify the conditional pdf for X given $\mathcal{G}$, and evaluate $\mathsf{E}[g(X) \mid \mathcal{G}] = \int g(x)\, f_X(x \mid \mathcal{G})\, dx$.

The easiest way to identify the conditional expectations posed above is to exploit independence, and to treat any variable on which we are conditioning as a constant. For example, since $(N_t - N_s) \perp\!\!\!\perp \mathcal{F}_s$ for $s < t$,

$$\begin{aligned}
\mathsf{E}[N_t \mid \mathcal{F}_s] &= \mathsf{E}[N_t - N_s \mid \mathcal{F}_s] + \mathsf{E}[N_s \mid \mathcal{F}_s] \\
&= \lambda(t-s) + N_s; \\
\mathsf{E}[e^{i\omega N_t} \mid \mathcal{F}_s] &= \mathsf{E}[e^{i\omega(N_t - N_s)} e^{i\omega N_s} \mid \mathcal{F}_s] \\
&= \mathsf{E}[e^{i\omega(N_t - N_s)}] e^{i\omega N_s} \\
&= \exp\big(\lambda(t-s)(e^{i\omega} - 1) + i\omega N_s\big); \\
\mathsf{E}[{N_t}^2 \mid \mathcal{F}_s] &= \mathsf{E}[(N_t - N_s)^2 + 2(N_t - N_s)N_s + {N_s}^2 \mid \mathcal{F}_s] \\
&= \lambda(t-s) + [\lambda(t-s)]^2 + 2\lambda(t-s)N_s + {N_s}^2
\end{aligned}$$

Notice that $\mathcal{F}_t := \sigma\{N_s : \ s \le t\}$ is a σ-algebra generated by *uncountably many* random variables N_s, but nevertheless we now are able to compute conditional expectations and probabilities given $\mathcal{F}_t$.

10.4 Borel's Paradox

Let (X,Y) be the longitude, $0 \le X < 2\pi$, and latitude, $-\pi/2 \le Y \le \pi/2$, of a point drawn uniformly from a sphere $\mathcal{S}$ (perhaps the globe). What is its *conditional* distribution of (X,Y), given that it lies on a great circle $\mathcal{C}$? This famously ill-posed question helps motivate a careful consideration of conditioning. If the "great circle" is the equator $Y = 0$, the answer is the (perhaps expected) uniform distribution, with longitude $X \sim \mathsf{Un}\big([0,2\pi)\big)$. But if the great circle is, say, the prime meridian $X = 0$, then the point is much more likely to be near the equator (where an interval of $Y = 0 \pm 1$ degree latitude has a large area) than near either pole (where it doesn't); in that case the conditional distribution of Y has density $f(y \mid x) = \frac{1}{2}\cos(y)\mathbf{1}_{[-\pi/2,\pi/2]}(y)$ for any $0 \le x < 2\pi$.

We simply cannot condition meaningfully on the null event that (X,Y) lies on a set of zero probability, such as a great circle. We *can* condition on events of positive probability, or on the σ-algebra generated by a random variable.

In *Radon spaces* (which include $\mathbb{R}^d$ and all complete separable metric spaces) these notions are closely related: in particular, we can always compute a version of the conditional expectation of one random-variable X given another Z as $\mathsf{E}[X \mid Z] = \phi_X(z)$ for the limit

$$\phi_X(z) = \limsup_{\epsilon \to 0} \mathsf{E}\big[\ X \mid \ \{|Z - z| < \epsilon\}\ \big].$$

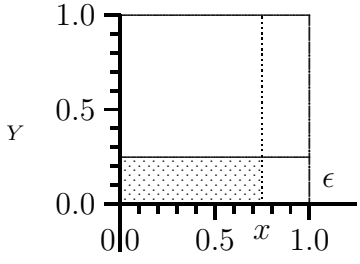
Let's use this to try to answer the question: *What is the conditional distribution of the horizontal component X of a point drawn from the unit square, given that the point lies on the bottom edge?* Let (X,Y) be the coordinates of a point drawn uniformly from the unit square and $0 < \epsilon < 1$, and let Δ denote the bottom edge of the square. For $0 < x < 1$ we can compute

$$\mathsf{P}[X \le x \mid 0 \le Y \le \epsilon] = \frac{\epsilon x}{\epsilon} = x$$

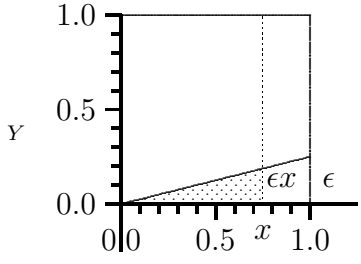
and conclude (taking $\epsilon \to 0$) that the conditional *distribution* of X, given $Y = 0$, is the standard uniform, and hence the conditional expectation $\mathsf{E}[X \mid Y = 0] = 1/2$. Similarly if we let $R := Y/X$ be the ratio of Y to X, we can also compute

$$\mathsf{P}[X \le x \mid 0 \le R \le \epsilon] = \frac{\epsilon x^2/2}{\epsilon/2} = x^2,$$

so the conditional distribution of X, given $R = 0$, is $\mathsf{Be}(2,1)$, with conditional density $f(x \mid R{=}0) = 2x$ on $[0,1]$ and conditional mean $\mathsf{E}[X \mid R{=}0] = 2/3$.



$\mathsf{P}[X \le x \mid Y \le \epsilon] = x$



$\mathsf{P}[X \le x \mid R \le \epsilon] = x^2$

Note that both of these "events" on which we condition are identical— the null event that (X,Y) lies on the bottom edge $\Delta := \{(x,0):\ 0<x\le 1\}$ of the square, another example of Borel's paradox. Really these two different results were answers to different questions: one found the values of $\mathsf{P}[X\le x \mid \sigma\{Y\}]$ and $\mathsf{E}[X \mid \sigma\{Y\}]$, the other found $\mathsf{P}[X\le x \mid \sigma\{R\}]$ and $\mathsf{E}[X \mid \sigma\{R\}]$. Geometrically, what do events in $\sigma\{Y\}$ and those in $\sigma\{R\}$ look like in the square? For an arbitrary density $f(x)$ on the unit interval, can you find a random variable Z (a function of X and Y) such that $\{Z=0\}$ is the bottom edge of the square and the conditional distribution of X given $Z=0$ is $f(x)\,dx$? Are any conditions on $f(x)$ needed?

A little more generally...

Let $f(x)$ be any strictly-positive bounded pdf on the unit interval, and set $Z := Y/f(X)$. Then $(X,Y)\in\Delta := [0,1]\times\{0\}$ if and only if $Z=0$ and, for $\epsilon>0$,

$$\begin{aligned}\mathsf{P}[X\le x \mid Z\le\epsilon] &= \frac{\mathsf{P}[X\le x\ \cap\ Y\le \epsilon f(X)]}{\mathsf{P}[Y\le\epsilon f(X)]}\\ &= \frac{\int_0^x \min[1,\epsilon f(\xi)]d\xi}{\int_0^1 \min[1,\epsilon f(\xi)]d\xi} = \frac{\int_0^x \min[1/\epsilon, f(\xi)]d\xi}{\int_0^1 \min[1/\epsilon, f(\xi)]d\xi}\\ &\to \int_0^x f(\xi)d\xi\end{aligned}$$

as $\epsilon\to 0$ by LMCT, so the limiting distribution of X conditional on $Z\le\epsilon$ is the completely arbitrary distribution with density $f(x)$. Thus, in a very strong way, the "conditional distribution of X given that $(X,Y)\in\Delta$" is not determined. We can find conditional probabilities and distributions given *random variables* or non-null *events* or (more generally than either) *sigma algebras*, but not given events of probability zero.

Be careful out there...

Borel's paradox isn't just an academic puzzle. Naïve attempts to "condition" on null events (for example, by trying to impose Bayesian prior distributions on both the inputs and outputs of deterministic models, as in *Inference from a Deterministic Population Dynamics Model for Bowhead Whales* by Raftery, Givens & Zeh, JASA 1995) pop up every year or two in the literature, and sometimes aren't caught in the review process. That one (I kid you not) led to discussions about Borel's Paradox at meetings of the International Whaling Commission, and in the 1995 IWC Annual Report (try googling "bowhead whale borel paradox").

Be careful!