Conditional expectation as Radon–Nikodym derivative

Alex Fu

2023-01-08

In the measure-theoretic formulation of probability theory, probability spaces are simply *measure spaces* with total measure 1; random variables are *measurable functions* from probability spaces; distributions are pushforward *measures* induced by random variables on their target space; and expectation is *Lebesgue integration* with respect to a probability measure.

But, much of probability and statistics is inferential rather than merely descriptive, and randomness with *given information* is often of interest. *Information*, in the informal sense, is characterized by events and σ*-algebras* in our formalization. **Conditioning** on given information can be more involved to formalize than unconditional probability, e.g. regular conditional probabilities, Markov transition kernels, random measures, conditional independence, Bayesian networks, etc. In this note, we will look at a definition of *conditional expectation* that requires some formal machinery, and in the process examine some reasons why this definition is more intuitive than it looks.

First, a bit about integration. The Lebesgue integral of f with respect to a measure μ is an *accumulation* of the values of f, just like Riemann integration; informally, it is a "weighted sum" of different values of y weighted by the mass $\mu(\{x : f(x) = y\})$. Addition is a special case of integration, being with respect to the counting measure. Expectation is a special case of integration, where $\mathbb{P}(\Omega) = 1$ ensures desirable properties such as the monotonicity of moments $\mathbb{E}(X^p)$ or the expected value of a constant being itself. We also have the following interpretation of integration:

> Integrals are a form of measures.

Think of familiar applications of Riemann integrals in calculus: finding the length of a parametric curve; surface area; volume; or the mass of an object with varying density. While Riemann integration is taken with respect to the usual uniform Lebesgue measure, physical density acts like a scaling factor that varies over space, or a "change-of-measure" factor. That is, if $m(R)$ measures the mass of the region R, then

$$m(R) = \int_R \rho(x) \, \mathrm{d}V(x).$$

Physical density ρ is the "relative rate" of the mass measure m with respect to the volume measure V; we could even write $\rho = \frac{\mathrm{d}m}{\mathrm{d}V}$ to reflect this relationship notationally, even if we have not yet defined the

derivative of a measure. However, taking inspiration from the analogy of physical density, we can define the *Radon–Nikodym derivative* of a measure with respect to another measure. In place of differentiability, the condition that a function is "locally linear" on an infinitesimal scale, or scales at most a finite factor, is the similar condition of *absolute continuity*, that one measure is "small" with respect to another, or at most finitely larger, never "infinitely larger."

Theorem 1 (Lebesgue decomposition theorem; Radon–Nikodym theorem)**.**

Let μ, ν be two σ-finite measures, which are in some sense "small" and uniquely determinable. Then there exist unique (σ-finite) measures $\mu^{\ll}, \mu^{\perp}$ such that $\mu = \mu^{\ll} + \mu^{\perp}$ and

- $\mu^{\ll}$ is **absolutely continuous** with respect to ν, denoted $\mu^{\ll} \ll \nu$: for every $A \in \Sigma$, $\nu(A) = 0$ implies $\mu^{\ll}(A) = 0$ (though not necessarily the converse);
- $\mu^{\perp}$ and ν are mutually **singular**, denoted $\mu^{\perp} \perp \nu$: there exists $B \in \Sigma$ such that $\mu^{\perp}(B^c) = 0$ and $\nu(B) = 0$. In other words, $\mu^{\perp}$ places all of its mass on B, and ν all of its mass on B^c.

Moreover, there exists a nonnegative measurable function f, unique ν-almost everywhere, such that

$$\mu^{\ll}(A) = \int_A f \, \mathrm{d}\nu$$

for every $A \in \Sigma$. We write $f := \frac{\mathrm{d}\mu^{\ll}}{\mathrm{d}\nu}$ for the **Radon–Nikodym derivative** of $\mu^{\ll}$ with respect to ν. We may further decompose $\mu^{\perp}$ as a sum of a *singular continuous* measure and a *discrete* measure, which we will not do here.

Intuitively, why does the absolute continuity of $\mu \ll \nu$ guarantee the existence of a density? There is no measurable set A with $\nu(A) = 0$ and $\mu(A) > 0$, which would require f to be infinite for

$$\mu(A) = \int_A f \, \mathrm{d}\nu$$

to hold in some sense. In other words, $\frac{\mu}{\nu}$ is never infinite, which ensures $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} < \infty$ as well. The uniqueness of the Lebesgue decomposition $\mu = \mu^{\ll} + \mu^{\perp}$ can be shown using the following result (absolute continuity and singularity are "orthogonal" properties).

Proposition 1.

If $\mu \ll \nu$ and $\mu \perp \nu$, then $\mu = 0$.

Let $\mathbb{P}_0, \mathbb{P}_1$ be two probability measures on (Ω, Σ), which are certainly (σ-)finite. Suppose that $\mathbb{P}_0 \ll \mathbb{P}_1$ and $\mathbb{P}_1 \ll \mathbb{P}_0$. Then the Radon–Nikodym derivative $\Lambda := \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}$ is the **likelihood ratio** of Neyman–Pearson hypothesis testing, where

$$\mathbb{P}_1(A) = \mathbb{E}_1(\mathbb{1}_A) = \int_A 1 \, \mathrm{d}\mathbb{P}_1 = \int_A \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0} \, \mathrm{d}\mathbb{P}_0 = \mathbb{E}_0(\Lambda \cdot \mathbb{1}_A)$$

for every event $A \in \Sigma$. As a measurable function (i.e. random variable), Λ describes the "relative density" of $\mathbb{P}_1$-mass relative to $\mathbb{P}_0$-mass for each sample point. It turns out that for the binary hypothesis testing

problem, the naïve solution of checking whether Λ is above a certain threshold λ is the (unique) optimal solution in the Neyman–Pearson sense.

The Radon–Nikodym derivative has another important application in probability: conditional expectation can precisely be formulated as a Radon–Nikodym derivative, which even gives the existence and almost sure-uniqueness of conditional expectation by Theorem 1.

Definition 1 (Conditional expectation).

Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with X integrable, i.e. $\mathbb{E}\,|X| < \infty$, and let $\mathcal{G} = \sigma(Y)$ be the σ-algebra generated by Y:

$$\sigma(Y) \coloneqq \left\{Y^{-1}(B) : B \in \mathcal{B}\right\} \subseteq \mathcal{F}.$$

We also define the measure μ on $\mathcal{F}$ as follows, which we can easily check defines a measure:

$$\mu(A) \coloneqq \int_A X \,\mathrm{d}\mathbb{P} = \mathbb{E}(X \cdot \mathbb{1}_A)$$

Then $Z \coloneqq \mathbb{E}(X \mid Y)$ is the **conditional expectation** of X given Y iff Z is $\mathcal{G}$-measurable, and

$$\int_A X \,\mathrm{d}\mathbb{P} = \int_A \mathbb{E}(X \mid Y) \,\mathrm{d}\mathbb{P}$$

for all $A \in \mathcal{G}$. Equivalently, $\mathbb{E}(X \mid Y)$ equals the Radon–Nikodym derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}} \restriction_{\mathcal{G}}$ almost surely.

The condition that $\mathbb{E}(X \mid Y)$ is $\sigma(Y)$-measurable becomes very natural with the following lemma, along with the intuition in the simpler discrete case that $\mathbb{E}(X \mid Y = y)$ is a function of y:

Lemma 1 (Measurability with respect to the generated σ-algebra).

A random variable Y is $\sigma(X)$-measurable iff there exists a measurable function $f \colon \mathbb{R} \to \mathbb{R}$ such that $Y = f(X) = f \circ X$.

For example, if Y only takes values in $1, \ldots, n$, then $\mathbb{E}(X \mid Y)$ is a random variable that is constant on each of the events $\{Y = 1\}, \ldots, \{Y = n\}$, and its value on $\{Y = y\}$ is the scaling factor c such that

$$\mu(\{Y = y\}) = \mathbb{E}(X \cdot \mathbb{1}_{Y=y}) = \int_{\{Y=y\}} X \,\mathrm{d}\mathbb{P} = \int_{\{Y=y\}} c \,\mathrm{d}\mathbb{P} = c \cdot \mathbb{P}(\{Y = y\}).$$

On average, X equals c on $\{Y = y\}$, which is what we expect from conditional expectation.

Physical density, probability density functions, likelihood ratios, conditional expectation, and Radon–Nikodym derivatives are functions that describe pointwise the "relative rate" between two measures.

■