Probability and Statistical Inference

Tianqi Zhang
Emory University

Apr 17th 2025

Theorem 0.0.1 Radon-Nikodym Theorem
Let $(X, \mathcal{M})$ be a measurable space, and let ν and μ be σ-finite measures on $(X, \mathcal{M})$. Then there exists a unique decomposition

$$\nu = \nu_a + \nu_s$$

such that:

1. $\nu_a \ll \mu$ (absolutely continuous part),
2. $\nu_s \perp \mu$ (singular part).

Moreover, there exists a unique (up to μ-a.e. equivalence) measurable function $f \geq 0$ such that $d\nu_a = f\, d\mu$. We denote this density function as the Radon-Nikodym derivative:

$$f = \frac{d\nu}{d\mu}$$

To prove this, we firstly need Riesz representation theorem for Hilbert Space:

Theorem 0.0.2 Riesz Representation Theorem (Hilbert Space)
Given $\mathcal{H}$ and a continuous linear functional

$$\ell : \mathcal{H} \to \mathbb{C}$$

Then $\exists!$ a vector $w \in \mathcal{H}$ such that $\forall v \in \mathcal{H}$,

$$\ell(v) = \langle v | w \rangle$$

Proof 1. We start by a proposition as follows:

Proposition 0.0.3 :
For a closed subspace $S \subset \mathcal{H}$, we have a decomposition

$$\mathcal{H} = S \oplus S^{\perp}$$

Such that $S^{\perp} \equiv \{w \in \mathcal{H} : \langle v|w\rangle = 0,\ v \in S\}$. Given $u \in \mathcal{H}$, set

$$d(u, S) \equiv \inf\{\|u - v\| : v \in S\}$$

Then there exists a sequence $\{v_n\}$ such that $\|u - v_n\| \to d$.

By the law of parallelogram, for any v_n, v_m in the sequence,

$$\|v_n - v_m\|^2 = 2\|u - v_n\|^2 + 2\|u - v_m\|^2 - 4\left\|u - \frac{v_n + v_m}{2}\right\|^2$$
$$\lim_{n,m\to\infty} \|v_n - v_m\|^2 \leq 2d^2 + 2d^2 - 4d^2 \leq 0$$

Therefore $\{v_n\}$ is Cauchy. Since S is closed, the limit exists and is in S, we denote it by v.

Define $w \equiv u - v$. We claim $w \in S^{\perp}$. Indeed, for any $s \in S$, consider

$$\|u - s\|^2 \ = \ \|(v + w) - s\|^2 \ = \ \|(v - s) + w\|^2.$$

Since v is the chosen minimizer in S, the first-order condition for minimality gives $\langle w, s-v\rangle = 0$. In particular, taking $s = v$ yields $\langle w, v-v\rangle = 0$ which is trivial, but more importantly, by choosing s to approach v suitably, we obtain $\langle w, s\rangle = 0$ for any $s \in S$. Thus w is orthogonal to every vector in S, i.e. $w \in S^{\perp}$.

Back to Riesz: We are given a continuous and linear functional $\ell : \mathcal{H} \to \mathbb{C}$. Take any $S \subset \mathcal{H}$ as the kernel of ℓ, i.e

$$S \equiv \ker(\ell) = \{v \in \mathcal{H} : \ell(v) = 0\}$$

Then ℓ is linear implies that S is a subspace (by Kernel), and ℓ is continuous implies that S is closed (by pre-image of the kernel). Then we have a decomposition by the proposition above:

$$\mathcal{H} = S \oplus S^{\perp}$$

Note that it would be trivial if either S or $S^{\perp}$ is the empty set. Then we can choose $w = 0$ and the theorem is proven.

Therefore, suppose that neither is the empty set. We fix any $w \in S^{\perp}$ with $\|w_1\| = 1$. We take $v \in \mathcal{H}$ and observe $v\ell(w_1) - w_1\ell(v) \in S$ since S is closed. Then by orthogonality,

$$\langle v\ell(w_1) - w_1\ell(v)|w\rangle = 0$$

We extend the expression into:

$$\left\langle v \middle| \overline{\ell(w_1)}w_1 \right\rangle - \ell(v)\|w_1\|^2 = 0$$

And finally:

$$\ell(v) = \left\langle v \middle| \overline{\ell(w_1)}w_1 \right\rangle$$

We define $w \in S^{\perp}$ to be $w \equiv \overline{\ell(w_1)}w_1$. We have shown the existence of $w \in S^{\perp}$ in this context.

As for uniqueness, assume $w' \in S^{\perp}$ that also suffices the above assumptions, then for all $v \in \mathcal{H}$.

$$\ell(v) - \ell(v) = \langle v|w\rangle - \langle v|w'\rangle$$
$$0 = \langle v|w - w'\rangle$$

We choose $v = w - w'$, then $\|w - w'\|^2 = 0$. By positive definite of the norm, we have $w = w'$. We have shown the uniqueness of such w.

Back to R-N: Let $(X, \mathcal{M})$ be a measure space with σ-finite measure μ, ν. Let $\rho = \mu + \nu$. Since both are sigma finite, ρ is also a properly defined measure on X. We define $\ell : \mathcal{L}^2(X, d\rho) \to \mathbb{C}$ by

$$\ell(\psi) \equiv \int_X \psi \, d\nu$$

Then since $\nu \leq \rho$,

$$|\ell(\psi)| = \langle 1|\psi\rangle_{\mathcal{L}^2(X,d\rho)} \leq \underbrace{\|1\|}_{<\infty} \|\psi\|_{\mathcal{L}^2(X,d\rho)} \leq C\|\psi\|_{\mathcal{L}^2(X,d\rho)}$$

Then ℓ is continuous.

By Riesz, $\exists g \in \mathcal{L}^2(X, d\rho)$ such that $\ell(\psi) = \langle\psi|g\rangle = \int_X \psi\overline{g} \, d\rho$. In particular for any $E \subset X$, we can express its measure

$$\nu(E) = \ell(\chi_E) = \int_E \overline{g} \, d\rho$$

Then g is a real and non-negative function a.e. on ρ.

Also, $\nu(E) \leq \rho(E)$, we have

$$\nu(E) = \int_E d\nu \leq \int_E d\rho \quad \forall E \in \mathcal{M}$$
$$\int_E g\, d\rho \leq \int_E d\rho$$

We have $g \leq 1$ a.e. on ρ.

Now given $\psi \in \mathcal{L}^2(X, d\rho)$.

$$\ell(\psi) = \int_X \psi g d\rho$$
$$\int_X \psi\, d\nu = \int_X \psi g\, d\mu + \int_X \psi g\, d\nu$$
$$\int_X \psi(1-g)\, d\nu = \int_X \psi g\, d\mu \tag{1}$$

As $g : X \to [0,1]$, we can define the set $A \equiv \{g < 1\}$ and $B \equiv \{g = 1\}$. We have accordingly the two measures $\nu_a(E) \equiv \nu(A \cap E)$ and $\nu_s(E) \equiv \nu(B \cap E)$. By (1) we have

$$\begin{aligned} \mu(B) &= \int_B d\mu \\ &= \int_X \chi_B g\, d\mu \quad \text{as } g = 1 \text{ on } B \\ &= \int_X \chi_B (1-g)\, d\nu = 0 \end{aligned}$$

Therefore, the measure μ cannot see B. We have $\nu_s \perp \mu$ proving the first half of the theorem.

Revisit (1), let $\psi \equiv \chi_E(1 + g + g^2 + \cdots + g^n)$ for any $E \in \mathcal{M}$. (1) equals to the following:

$$\int_E \left(1 - g^{n+1}\right) d\nu = \int_E g(1 + \cdots + g^n)\, d\mu$$
$$\int_{E\cap A} 1 - g^{n+1}\, d\nu = \int_E g(1 + \cdots + g^n)\, d\mu$$

If we take $n \to \infty$,

The left hand side will be dominated by $\nu_a(E)$ since $g < 1$ on A. Then by DCT, we have the left as $\int_E d\nu_a$.

The right hand side will converge to $\int_E \frac{g}{1-g}\, d\mu$ as a geometric series. We have the following expression:

$$\nu_a(E) = \int_E \frac{g}{1-g}\, d\mu$$

We define $f \equiv \frac{g}{1-g}$. Therefore, we have the Radon-Nikodym derivative.

■

Remark 0.1 *Conditional Expectation:*
In machine learning, the fundamental problem is a minimization of the square error:

$$h^* = \arg\min_{h\in\mathcal{H}} \mathbb{E}\big[(Y - h(X))^2\big]$$

where: $\mathcal{H}$ is a functional space (often Hilbert space). Y is the target or output, X is the input, and the expectation is over the joint distribution of (X, Y) induced by some measure P on the product space $(\Omega, \mathcal{F})$. The solution of such problem is uniquely given by

$$h^*(X) = \mathbb{E}[Y|X]$$

The Radon-Nikodym theorem underlies the modern theory of conditional expectation, a foundational concept in machine learning.

Given data X sampled from a probability space $(\Omega, \mathcal{F}, P)$, the observed information induces a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, often written as $\mathcal{G} = \sigma(X)$. The conditional expectation of an integrable random variable Y given $\mathcal{G}$ is defined as the unique $\mathcal{G}$-measurable function $\mathbb{E}[Y \mid \mathcal{G}] : \Omega \to \mathbb{R}$ such that

$$\int_A \mathbb{E}[Y \mid \mathcal{G}]\, dP = \int_A Y\, dP \quad \textit{for all } A \in \mathcal{G}.$$

This implies that the function $\mathbb{E}[Y \mid \mathcal{G}]$ is the unique (up to P-a.e.) Radon-Nikodym derivative of the signed measure $\nu(A) = \int_A Y\, dP$ with respect to $P|_{\mathcal{G}}$, the restriction of P to $\mathcal{G}$. Thus, conditional expectation emerges naturally as a Radon-Nikodym derivative and represents the best $\mathcal{G}$-measurable approximation to Y.