Statistics 212: Lecture 2 (January 29th, 2025)

Conditional Expectation and Martingales

Instructor: Mark Sellke

Scribes: Hasan Laith and Shirley Xiong

1 January 29: Conditional Expectation and Martingales

1.1 Administration

Poll for SC 705 v.s. Sever 103 at end of class. (Sever won.)

1.2 Finishing Up Radon-Nikodym

Recall the *Radon-Nikodym Theorem*, where we have finite measures μ and ν. If $\nu \ll \mu$, (i.e. ν is absolutely continuous with respect to μ, or that $\mu(S) = 0 \implies \nu(S) = 0$) then there exists integrable non-negative f s.t. $f = \frac{d\nu}{d\mu}$ (call this the RN derivative) and (where $\omega \in \Omega$):

$$\nu(S) = \int_S f(\omega)\, d\mu(\omega) = \int_S f\, d\mu$$

We have these three variants:

(a) There exists an integrable, non-negative function f and another finite measure Θ such that:

$$\nu(S) = \Theta(S) + \int_S f(\omega)\, d\mu(\omega)$$

and Θ, μ are disjointly supported, meaning that there exists $S \in \mathscr{F} : \mu(S) = 0$, and $\Theta(\Omega \setminus S) = 0$. So, we are decomposing our ν into the absolutely continuous part (the integral) and the singular part (Θ) (we can decompose the singular part further .

(b) Let ν, μ be finite measures on $(\Omega, \mathscr{F})$, such that $0 \le \nu(S) \le \mu(S)$ for all $S \in \mathscr{F}$. Then:

$$\exists\, f : \Omega \to [0,1], \text{ and } f = \frac{d\nu}{d\mu}$$

That is for all $S \in \mathscr{F}$:

$$\int_S f(\omega)\, d\mu(\omega) = \nu(S).$$

(c) Let $\nu \ll \mu$ be probability measures. Then for all S:

$$\exists\, f : \Omega \to \mathbb{R}_{\ge 0} \text{ which we denote } f = \frac{d\nu}{d\mu},\ \int_S f\, d\mu = \nu(S) \text{ and } \int_\Omega f\, d\mu = 1 = \mathbb{E}^M[f].$$

The steps for (3):
The general idea here is to exhaust from below. Initially define the set:

$$H = \left\{ \text{measurable functions } f : \int_S f\,d\mu \le \nu(S)\ (\forall S \in \mathscr{F}) \right\}$$

(where $f : \Omega \to [0,1]$). Then, we want to find $f_* \in H$, with $\int_\Omega f_*\,d\mu = M \equiv \sup_{f\in H}\int_\Omega f\,d\mu$.

We want $f_* = \frac{d\mu}{d\nu}$ and show that it is the maximal element (s.t. we then have the desired equality with $\nu(S)$ for all S) it remains to assume that there exists $S \in \mathscr{F}$ such that $\int_S f_*\,d\mu \le \nu(S)$ and get a contradiction as if the increased element is contained by H, then f_* wouldn't be maximal.

First, we remove $E_1 = \{\omega : f_*(\omega) = 1\}$; for all $S' \subseteq E_1 : \mu(S') = \nu(S')$, since $\mu \ge \nu$ by assumption and $\mu(S') = \int_{S'} f_*\,d\mu \le \nu(S'), f_* \in H$.

Now, we can assume that $S \subseteq E_0 = \Omega \backslash E_1$, and write $\bar{S} \equiv S \cap F_n$ for notational convenience.

Using $\int_S f_*\,d\mu \le \nu(S)$, we know that $\exists n > 0$, where the same holds for $F_n = \{\omega : f_*(\omega) \le 1 - 1/n\}$, as:

$$F_1 \subseteq F_2 \subseteq \cdots \subseteq \cdots \to E_0 \text{ by Monotone Convergence Theorem.}$$

Then:

$$\int_{\bar{S}} f_*\,d\mu < \nu(\bar{S}) \implies \mu(\bar{S}) > \nu(\bar{S}) > 0.$$

So now our attempt is to hope that $f_* + \epsilon\chi_{\bar{S}} \in H$ for some small $\varepsilon > 0$, which would increase maximality of M since we'd be increasing f_*. This might look sufficient since:

- if $\varepsilon < 1/n$, then $f_* + \epsilon\chi_{\bar{S}} : \Omega \to [0,1]$.
- $\int_{\bar{S}} f_* + \varepsilon\chi_{\bar{S}} d\mu < \nu(\bar{S})$

However, the condition for H (by construction) needs to hold for all sets in $\mathscr{F}$, and not just $\bar{S}$ (and those which are then contained in $\bar{S}$). *So the idea is to prune down $\bar{S}$ in a way that fixes everything, removing violators* — once we get rid of all the violating sets, then we'll be happier.

Definition: Allow the *deficit* to be $\mathrm{Def}_\varepsilon = \nu(A) - \int_A f_*\,d\mu - \varepsilon\mu(A)$. This is countably additive for disjoint sets A, B. (Think of this as an error term of sorts.)

We've just seen that $\mathrm{Def}_\varepsilon(\bar{S}) > 0$. Let's suppose that some $S_1 \subseteq \bar{S}$ violates the condition to be in H for $f_* + \varepsilon\chi_{\bar{S}}$, which exactly says that $\mathrm{Def}_\varepsilon < 0$, implying that the deficit increases if we just remove $S_1 : \mathrm{Def}_\varepsilon(\bar{S}\backslash S_1) > \mathrm{Def}_\varepsilon(\bar{S}) > 0$.

Define disjoint $S_1, S_2, \cdots \subseteq \bar{S}$, which each optimize the deficit leftover. Recursively, define

$$a_k = \inf_{S_k \le \bar{S} \text{ disjoint from previous} S_1,\ldots,S_{k-1}} \mathrm{Def}_\varepsilon(S_k) \le 0.$$

In each step, we choose S_k to ensure that $\mathrm{Def}_\varepsilon(S_k) \le a_k + \frac{1}{k}$, and let $\hat{S} = \bar{S}\backslash(S_1 \cup S_1 \cdots \cup \ldots)$.

We claim that $f_* + \varepsilon\chi_{\hat{S}} \in H$. It follows from the definitions about deficit that $\mathrm{Def}_\varepsilon(\hat{S}) > 0 \implies \nu(\hat{S}) > 0 \implies \mu(\hat{S}) > 0$, so we actually get a contradiction since we've increase f_* by a positive amount.

Suppose that $\exists\tilde{S} \subseteq \hat{S} : \mathrm{Def}_\varepsilon(\tilde{S}) < -1/k < 0$. This causes a contradiction since we should have used $S_k \cup \tilde{S}$ instead of S_k, i.e. S_k disobeys its definition since it isn't within $1/k$ within minimizing the deficit and thus we haven't exhausted everything we've meant to be exhausting.

Note: sets disjoint from $\bar{S}$ don't matter. This is because we already had $f_* \in H$, so the only potential violations of the condition to be in H come from sets within $\bar{S}$. (I.e. for general $A \in \mathscr{F}$, check separately for $A \cap \bar{S}$ and $A\backslash\bar{S}$ and add.)

1.3 Conditional Expectation

Definition: Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and $\mathscr{G} \subseteq \mathscr{F}$, $X \in L^1(\mathbb{P}, \mathscr{F})$. Note:

- $X \in L^1(\mu) \implies \mathbb{E}^M|X| < \infty$
- $\mathscr{G} \subseteq \mathscr{F} \iff G$ is a sub $\sigma-$algebra/σ-field $\iff S \in \mathscr{G} \implies S \in \mathscr{F}$

Then, $\mathbb{E}(X|\mathscr{G})$ is the unique (up to measure 0) $\mathscr{G}-$measurable function such that:

$$\mathbb{E}(X \cdot \chi_S) = \mathbb{E}\left[\mathbb{E}[X|\mathscr{G}] \cdot \chi_S\right] \forall S \in \mathscr{G}$$

That is, we should treat the *conditional expectation* $\mathbb{E}[X|\mathscr{G}]$ as a random variable with the above property, and in terms of integrals (recalling that multiplying by an indicator translates to integrating over its support), for all $S \in \mathscr{G}$:

$$\int_S X \, d\mathbb{P} = \int_S (\mathbb{E}[X|\mathscr{G}]) \, d\mathbb{P}$$

Connecting back to what we might have seen before for conditional expectation, *for a random variable Y:*

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$

(where $\sigma(Y)$ is the smallest $\sigma-$algebra such that Y is measurable).

1.3.1 Existence and Uniqueness

Uniqueness:

Suppose that $Y = \mathbb{E}(X|\mathscr{G}) = Z$. Let $S_+ = \{Y > Z\}, S_- = \{Y < Z\}$. Note that both of these have to have measure 0, since they are both $\mathscr{G}-$measurable ($S_+, S_- \in \mathscr{G}$).

That is:

$$\mathbb{E}\left[(Y - Z) \cdot \chi_{S_+}\right] = \mathbb{E}\left[(X - X) \cdot \chi_{S_+}\right]$$

And clearly the right-hand term evaluates to 0 so specifically $Y = Z$ a.s. in order for the above to hold; we have the same idea for S_-.

We can also take this in terms of integrals and etc.

The proof for the uniqueness of the RN result is similarly as so:

If $f_1, f_2 = \frac{dx}{d\mu}$, and construct the sets $S_+ = \{f_1 > f_2\}, S_- = \{f_1 < f_2\}$ then naturally they have measure 0 a.s.. We would have, assuming if $\mu(S_+) > 0$:

$$\nu(S_+) - \nu(S_+) = \int_{S_+} (f_1 - f_2) d\mu > 0$$

Since S_+ is where $f_1 > f_2$ and a non-zero measure set then the integral must be positive; however, under RN the integral is equal to our left-hand difference (by linearity of integral which is a contradiction (and we can do the same for S_- with $f_2 - f_1$ in the integrand), so we are done.

Existence: We can prove this directly by the RN Theorem.

Assume that $X \geq 0$ (in general, decompose $X = X_+ - X_-$ and do the below argument for both). Let ν be a finite measure on $\mathscr{G}$ given by $\nu(S) = \int X d\mathbb{P}$ (where $\mathbb{P}$ is the probability measure on the original space). Then, $\nu \ll \mathbb{P}$, considering $\mathbb{P}$ as a $\mathscr{G}-$measure by restricting the sets plugged into $\mathbb{P}$. This implies that there exists:

$$\mathscr{G}\text{-measurable function which we write } \frac{d\nu}{d\mathbb{P}} = \mathbb{E}(X|\mathscr{G}).$$

This matches the definition of an RN derivative since, if $S \in \mathscr{G} : \nu(S) = \int_S \mathbb{E}(X|\mathscr{G}) d\mathbb{P}$ allegedly. By definition, this is equal to $\int_S X d\mu$, which is exactly equal to $\mathbb{E}[\mathbb{E}[X|\mathscr{G}] \cdot \chi_S]$, which is also equal to $\mathbb{E}[X\chi_S]$, and then the allegation follows by transitivity of equalities.

1.3.2 Practice with the Definition

+ **Showing that** $\mathbb{E}[aX_1 + bX_2|\mathcal{G}] = a\mathbb{E}[X_1|\mathcal{G}] + b\mathbb{E}[X_2|\mathcal{G}]$:

Given the separate conditional expectations $\mathbb{E}[X_1|\mathcal{G}], \mathbb{E}[X_2|\mathcal{G}]$ are $\mathcal{G}-$measurable, then the sum $a\mathbb{E}[X_1|\mathcal{G}] + b\mathbb{E}[X_2|\mathcal{G}]$ is also $\mathcal{G}-$measurable.

We want to show that if $S \in \mathcal{G}$, then $\int_S aX_1 + bX_2 d\mathbb{P} \stackrel{?}{=} \int_S a\mathbb{E}[X_1|\mathcal{G}] + b\mathbb{E}[X_2|\mathcal{G}] d\mathbb{P} \stackrel{?}{=} \mathbb{E}(aX_1 + bX_2|\mathcal{G})$. By linearity, $\int_S aX_1 d\mathbb{P} = \int_S a\mathbb{E}(X_1|\mathcal{G})$, and similarly for X_2, and then the proof is finished as long as we're not confused.

+ **For** $\mathcal{G} = \{\{\}, A, A^c, \Omega\}$, **check the conditional expectation.**

To say that a function is $\mathcal{G}$-measurable here, it just means ($\iff$) that it is constant in A, and constant in A^c.

Here, we'll only be taking expectation over A, A^c, which is just the average-taking process we are used to.

1.3.3 Conditional Expectations Decrease Convex Functions

Theorem (Jensen's Inequality): let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. Then,

$$\phi(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[\phi(X)|\mathcal{G}], \quad \mathcal{G} - a.e.$$

It is helpful to think of how to prove it in this case. Here, we want to show that if $S \in \mathcal{G}$, then:

$$\int_S \phi(\mathbb{E}[X|\mathcal{G}]) d\mathbb{P} \leq \int_S \mathbb{E}[\phi(X)|\mathcal{G}] d\mathbb{P}.$$

What is special about convex functions is that ϕ is convex if and only if for all $x \in \mathbb{R}$, there exists $\phi'(x) \in \mathbb{R}$ such that $\phi(Y) - \phi(X) \geq \phi'(x) \cdot (y - x)$ for all $y \in \mathbb{R}$ (visualize as basically saying that if you draw the tangent line to the graph of ϕ, it will stay below).

We're going to use that

$$\phi(X) \geq \phi(\mathbb{E}[X|\mathcal{G}]) + (X - \mathbb{E}[X|\mathcal{G}]) \cdot \phi'(\mathbb{E}[X|\mathcal{G}]),$$

where we're just drawing a tangent line at $\mathbb{E}[X|\mathcal{G}]$ and using convexity. Note also that

$$\int_S \mathbb{E}[\phi(X)|\mathcal{G}] d\mathbb{P} = \int_S \phi(X) d\mathbb{P}.$$

Then, rearranging, it suffices to show that $\int_S (X - \mathbb{E}[X|\mathcal{G}]) \cdot \phi'(\mathbb{E}[X|\mathcal{G}]) = 0$. The difference in the integrand vanishes when hit with $\mathcal{G}-$measurable stuff. The standard measure-theoretic argument is that we approximate $\phi'(\mathbb{E}[X|\mathcal{G}])$ by simple functions which take finitely many values.

This has a couple of nice consequences, as follows.

1.3.4 Contraction in L^1

Corollary:

$$\phi(X) = |X| \implies |\mathbb{E}[X|\mathcal{G}]| \leq \mathbb{E}[|X| \,|\, \mathcal{G}].$$

If we integrate the both sides, we'll see that:

$$\|\mathbb{E}[X|\mathscr{G}]\|_{L^1} \leq \|X\|_{L^1}, \text{ since } \int_\Omega \mathbb{E}[|X||\mathscr{G}]d\mathbb{P} = \int_\Omega |X|d\mathbb{P} \text{ as } \omega \in \mathscr{G}.$$

Corollary: If $X_1, \ldots, X_n \xrightarrow{L_1} X$, (i.e. $\|X - X_n\|_{L^1} \to 0$), then:

$$\int \mathbb{E}[|X - X_n||\mathscr{G}] \to 0 \implies \mathbb{E}[X_n|\mathscr{G}] \xrightarrow{L_1} \mathbb{E}[X|\mathscr{G}].$$

1.3.5 Projection in L^2

If $X \in L^2(\mathscr{F})$, then $\mathbb{E}[X|\mathscr{G}]$ is closest $L^2(\mathscr{F})$ approximation which is $\mathscr{G}-$measurable.
In fact, if $Y \in L^2(\mathscr{G})$, then $\mathbb{E}[(X - (\mathbb{E}[X|\mathscr{G}] + Y))^2] = \mathbb{E}[(X - \mathbb{E}[X|\mathscr{G}])^2] + \mathbb{E}[Y^2]$, so we really just have a Pythagorean theorem.

> *Proof*: We want cross term to vanish, meaning that $\mathbb{E}[(X - \mathbb{E}[X|\mathscr{G}]) \cdot Y] = 0$. To show this, write $Y = Y_+ - Y_-, Y \geq 0$ to assume WLOG that $Y \geq 0$. Then have a sequence of increasing approximations $Y_1 \leq Y_2 \leq \ldots \uparrow Y$, where each Y_n is a simple function. (For example, make Y_n the largest multiple of 2^{-n} which is smaller than Y and at most 2^n.) Then, by dominated convergence on their squares, we get that $\|Y_n\|_{L^2} \to \|Y\|_{L^2}$, which implies that $\|Y - Y_n\|_{L^2} \to 0$. To see the last implication, note that if $0 \leq Y' \leq Y$ almost surely, then
>
> $$\|Y - Y'\|^2_{L^2} = \mathbb{E}[(Y - Y')^2] \leq \mathbb{E}[Y^2] - \mathbb{E}[(Y')^2]$$
>
> since $a^2 + b^2 \leq (a + b)^2$ for any $a, b \geq 0$. Then we can say that $\mathbb{E}[(X - \mathbb{E}[X|\mathscr{G}]) \cdot (Y - Y_n)] \to 0$ by Cauchy-Schwarz. Meanwhile by again using the definition of conditional expectation and breaking Y_n into a sum of finitely many indicators, we have $\mathbb{E}[(X - \mathbb{E}[X|\mathscr{G}]) \cdot Y_n] = 0$ for all n. Combining finishes the proof.

1.4 Martingales

Definition: A *filtration* is a sequence of $\sigma-$algebra satisfying $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq F_2 \subset \ldots$ ((weakly) monotonically increasing).

Definition: A stochastic process $(X_t)_{t\geq 0}$ is *adapted* to $(\mathscr{F}_t)$ if X_t is $\mathscr{F}-$measurable.

Definition: A stochastic process $(X_t)_{t\geq 0}$ is a *martingale* if $X_t \in L^1$ and $\mathbb{E}[X_{t+1}|\mathscr{F}_t] = X_t$ (both for all t).
We can also say that $(X_t)_{t\geq 0}$ is also a martingale relative to the "natural filtration" if we have for all t $\mathscr{F}_t = \sigma(X_1, \ldots, X_t)$.