STAT 38100 Final Review

Rohan Buluswar

March 2025

Introduction

These are my final review notes for STAT 38100, Measure-Theoretic Probability I, at the University of Chicago with Prof. Chao Gao, in the Winter 2025 quarter.

1 Set Families

Definition 1.1. *Let $(A_n)_n$ be a sequence of sets. We define*

$$\limsup A_n = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$$

$$\liminf A_n = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n$$

If the two are the same, we define

$$\liminf A_n = \lim_{n\to\infty} A_n = \limsup A_n$$

Remark: We note that $x \in \liminf A_n \iff x$ is in all but finitely many of the A_n and $x \in \limsup A_n \iff x$ is in infinitely many of the A_n. Thus, $\liminf A_n \subseteq \limsup A_n$.

Remark: If $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$, i.e. the A_n are increasing, then

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$$

and if $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$, i.e. the A_n are decreasing, then

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$$

Now, given a universal set Ω, we examine some classes of subsets of $\mathcal{P}(\Omega)$, i.e. set families over Ω.

Definition 1.2 (σ-algebras). *Consider a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$. We say that $\mathcal{F}$ is a σ-algebra if it satisfies the following properties:*

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3. $A_n \in \mathcal{F} \implies \bigcup_n A_n \in \mathcal{F}$

Remark: If $\mathcal{F}_\lambda$ are σ-algebras for $\lambda \in \Lambda$, then $\bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ is also a σ-algebra.

Definition 1.3. *Consider any collection of subsets $\mathcal{E} \subseteq \mathcal{P}(\Omega)$. We define the σ-algebra generated by $\mathcal{E}$, $\sigma(\mathcal{E})$, to be the smallest σ-algebra containing $\mathcal{E}$. That is, $\mathcal{E} \subseteq \sigma(\mathcal{E})$ and given any σ-algebra $\mathcal{F} \supseteq \mathcal{E}$, we have $\mathcal{F} \supseteq \sigma(\mathcal{E})$.*

Remark: For a constructive definition, we can see that the intersection of all σ-algebras over Ω that contain $\mathcal{E}$ will satisfy the necessary properties.

Definition 1.4 (π-system). *A π-system is a collection of subsets $P \subseteq \mathcal{P}(\Omega)$ such that for any $A, B \in P$, $A \cap B \in P$.*

Definition 1.5 (Monotone class). *A monotone class is a collection of subsets $M \subseteq \mathcal{P}(\Omega)$ such that for any sequence $A_n \in M$, if $(A_n)_n$ is nested increasing or decreasing, we have $\lim_n A_n \in M$ as well.*

Definition 1.6 (λ-system). *A λ-system is a collection of subsets $\Lambda \subseteq \mathcal{P}(\Omega)$ such that*

1. $\Omega \in \Lambda$
2. *If $A, B \in \Lambda$ such that $A \supseteq B$, then $A \setminus B \in \Lambda$*
3. *If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ and each $A_n \in \Lambda$ then $\bigcup_n A_n \in \Lambda$*

Definition 1.7 (Algebra). *An algebra is a collection of subsets $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ such that*

1. $\emptyset \in \mathcal{A}$
2. $A \in \mathcal{A} \implies A^c \in \mathcal{A}$
3. $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$

Definition 1.8. *Given a set family $\mathcal{E} \subseteq \mathcal{P}(\Omega)$, we define $m(\mathcal{E})$ to be the smallest monotone class containing $\mathcal{E}$ and we define $\lambda(\mathcal{E})$ to be the smallest λ-system containing $\mathcal{E}$.*

Theorem 1.1 (**Monotone Class Theorem**). *If $\mathcal{A}$ is an algebra then $m(\mathcal{A}) = \sigma(\mathcal{A})$.*

Theorem 1.2 ($\pi - \lambda$ theorem). *If P is a π-system, then $\lambda(P) = \sigma(P)$.*

Definition 1.9 (**Measurable spaces**). *If Ω is a set and $\mathcal{F}$ is a σ-algebra over Ω, the pair $(\Omega, \mathcal{F})$ is called a measurable space.*

Example: We can take $\Omega = \mathbb{R}$ and $\mathcal{B}_{\mathbb{R}} = \sigma(\{U \subseteq \mathbb{R} : U \text{ is open}\}) = \sigma(\{[a, \infty) : a \in \mathbb{R}\})$. For another example, we can take $\Omega = [0, 1]$ and $\mathcal{B}_{[0,1]}$ to be generated by unions of finitely many intervals in $[0, 1]$.

2 Defining Measure, Extension Theorems

Definition 2.1 (**Measure**). *Consider a set family $\mathcal{E} \subseteq \mathcal{P}(\Omega)$. We say that a function $\mu : \mathcal{E} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ is a measure if it satisfies:*

1. $\mu(\emptyset) = 0$
2. *If $A_n \in \mathcal{E}$ are disjoint and $\bigcup_n A_n \in \mathcal{E}$ then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.*

Definition 2.2. *A measure μ is called finite if for all $A \in \mathcal{E}$, $\mu(A) < \infty$. It is instead called σ-finite if for all $A \in \mathcal{E}$, there exist $A_n \in \mathcal{E}$ such that $A \subseteq \bigcup_n A_n$ and $\mu(A \cap A_n) < \infty$ for all n.*

Together, the triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space. If $\mu(\Omega) = 1$, it is called a probability space, and μ is a probability.

Proposition 2.1.

1. *If $A \subseteq B$ then $\mu(A) \leq \mu(B)$*
2. *If $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ then $\mu(\lim_n A_n) = \lim_n \mu(A_n)$*
3. *If $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$ and $\mu(A_1) < \infty$ then $\mu(\lim_n A_n) = \lim_n \mu(A_n)$*

It may be easy to define a length measure on an algebra, such as the collection of unions of finitely many intervals in $[0, 1]$. However, it is not obvious if that measure can be extended uniquely to a σ-algebra, in order to have a proper measure space. To that end, we have the following theorems.

Theorem 2.1. *Suppose we have finite measures μ and ν on a σ-algebra $\sigma(\mathcal{A})$, such that $\mu \equiv \nu$ on $\mathcal{A}$. Then in fact $\mu \equiv \nu$ on all of $\sigma(\mathcal{A})$.*

(This theorem answers the question of uniqueness of extension.)

Definition 2.3 (**Outer measure**). *We call $\tau : \mathcal{P}(\Omega) \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ an outer measure if it satisfies the following:*

1. $\tau(\emptyset) = 0$
2. $A \subseteq B \implies \tau(A) \leq \tau(B)$
3. $\tau(\bigcup_n A_n) \leq \sum_n \tau(A_n)$

Theorem 2.2. *Suppose we have a set family $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ such that $\emptyset \in \mathcal{E}$, and let μ be a nonnegative function on $\mathcal{E}$ such that $\mu(\emptyset) = 0$. We can define*

$$\tau(A) = \inf\{\sum_n \mu(B_n) : \bigcup_n B_n \supseteq A, B_n \in \mathcal{E}\}$$

Then τ is an outer measure on $\mathcal{P}(\Omega)$.

Theorem 2.3 (**Caratheodory Extension Theorem**). *Let τ be an outer measure. We can restrict it to*

$$\mathcal{F}_\tau = \{A \subseteq \Omega : \tau(B) = \tau(B \cap A) + \tau(B \cap A^c), \forall B \subseteq \Omega\}$$

Then $(\Omega, \mathcal{F}_\tau, \tau)$ is a measure space.

We can combine these theorems to obtain the following result:

Theorem 2.4. *If μ is a finite measure on an algebra $\mathcal{A}$, then it can be uniquely extended to $\sigma(\mathcal{A})$.*

Now, we can construct the uniform probability measure on $[0,1]$. To begin, we define $\mathcal{A}$ to be the set of unions of finitely many intervals in $[0,1]$, with λ being the natural length measure on $\mathcal{A}$. We then extend λ to $\sigma(\mathcal{A}) = \mathcal{B}_{[0,1]}$. However, in the proof of the theorem, we actually extended it to $\mathcal{F}_\tau$. We know $\sigma(\mathcal{A}) \subseteq \mathcal{F}_\tau$, but it may be the case that the inclusion is strict. In this case, it is, and $\mathcal{F}_\tau$ is the Lebesgue σ-algebra, $\mathcal{L}_{[0,1]}$. To prove this fact, we introduce the following definition and refinement of an earlier result:

Definition 2.4 (Completeness of measure)**.** *A measure space $(\Omega, \mathcal{F}, \mu)$ is called complete if for any $A \in \mathcal{F}$ such that $\mu(A) = 0$ and for any $B \subseteq A$, $B \in \mathcal{F}$.*

Theorem 2.5 (Caratheodory Extension Theorem)**.** *Let τ be an outer measure and let us define*

$$\mathcal{F}_\tau = \{A \subseteq \Omega : \tau(B) = \tau(B \cap A) + \tau(B \cap A^c), \forall B \subseteq \Omega\}$$

Then $(\Omega, \mathcal{F}_\tau, \tau)$ is a complete measure space.

Proposition 2.2. $([0,1], \mathcal{B}_{[0,1]}, \lambda)$ *is not a complete measure space.*

(As a result of this proposition, we have $\mathcal{B}_{[0,1]} \subsetneq \mathcal{L}_{[0,1]}$.)

Theorem 2.6 (Completion of measure spaces)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We can define*

$$\tilde{\mathcal{F}} = \{A \cup N : A \in \mathcal{F}, N \subseteq B \in \mathcal{F}, \mu(B) = 0\}$$

$$\tilde{\mu}(A \cup N) = \mu(A)$$

Then $\tilde{\mu}$ is a well-defined measure on $\tilde{\mathcal{F}}$ and $(\Omega, \tilde{\mathcal{F}}, \tilde{\mu})$ is a complete measure space, which is called the completion of $(\Omega, \mathcal{F}, \mu)$. Moreover, $\mu \equiv \tilde{\mu}$ on $\mathcal{F}$.

Theorem 2.7. *Suppose μ is a finite measure on an algebra $\mathcal{A}$ and τ is the outer measure generated by μ and coverings, as above. Then $\mathcal{F}_\tau$ is the completion of $\sigma(\mathcal{A})$.*

Lemma 2.8. *In the above setting, for any $A \in \mathcal{F}_\tau$, there exists $B \in \sigma(A)$ such that $A \subseteq B$ and $\tau(B) = \tau(A)$.*

As a consequence of the above theorem, we know that $\mathcal{L}_{[0,1]}$ is the completion of $\mathcal{B}_{[0,1]}$

3 Lebesgue Integration

Definition 3.1 (Measurable functions). *A map $f : (\mathcal{X}, \mathcal{F}) \to (\mathcal{Y}, \mathcal{G})$ is called measurable if for all $G \in \mathcal{G}$, $f^{-1}(G) \in \mathcal{F}$, i.e. $f^{-1}\mathcal{G} \subseteq \mathcal{F}$.*

Definition 3.2 (Random variables). *A measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is called a random variable.*

Lemma 3.1. *Suppose we have a function $f : (\mathcal{X}, \mathcal{F}) \to (\mathcal{Y}, \sigma(\mathcal{E}))$ such that $f^{-1}(\mathcal{E}) \subseteq \mathcal{F}$. Then f is measurable.*

Now, we construct some examples of measurable functions. Note that given a function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$, to show that X is a random variable, it suffices to show that the set $\{\omega : X(\omega) > a\} \in \mathcal{F}$ for all $a \in \mathbb{R}$.(This is because the sets $\{x > a\} \subseteq \mathbb{R}$ generate the topology, and thus the Borel family of sets.) Sets of this kind are typically abbreviated as $\{X > a\} \subseteq \Omega$. We can instead check only for $a \in \mathbb{Q}$, and the condition $X > a$ can be replaced by $X \geq a$, among other choices.

Example 3.1. The indicator function $\mathbb{I}_A$ is measurable $\iff A \in \mathcal{F}$.

Definition 3.3 (**Simple function**)**.** *If $A_1, A_2, ..., A_n \in \mathcal{F}$ and the A_i form a partition of Ω, then for any coefficients $a_i \geq 0$, the function*

$$f = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i}$$

is called a simple function, and is measurable.

To get general (nonnegative) measurable functions, we can take the limit of approximations from below by simple functions. Specifically, we have the following:

Lemma 3.2. *Suppose $f \geq 0$ is a measurable function. Then there exists a sequence of nonnegative simple functions $(f_n)_n$ such that $f_n \nearrow f$ pointwise as $n \to \infty$. Moreover, if f is bounded, then the convergence is uniform.*

Proof. We can define

$$f_n = n\mathbb{I}_{\{f \geq n\}} + \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbb{I}_{\{\frac{k}{2^n} \leq f < \frac{k+1}{2^n}\}}$$

□

Finally, given a general measurable function f, we can write $f = f^+ - f^-$ where $f^+ = \max(f, 0)$ and $f^- = (-f)^+$. Thus, our construction was as follows: indicator functions $\implies$ simple functions $\implies$ nonnegative measurable functions $\implies$ general measurable functions. Our definition of the integral and proofs of various properties will follow the same route.

Proposition 3.1. *The following operations preserve measurability: sums, products, differences, supremum, infimum, max, min, and composition.*

Definition 3.4 (**Lebesgue integral**)**.** *Measure-theoretic integration is defined in four steps, as outlined before. Suppose we have a measurable function $f : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$. Then*

1. *If $f = \mathbb{I}_A$ for some $A \in \mathcal{F}$ then $\int f d\mu = \mu(A)$*
2. *If $A_1, ..., A_n$ form a partition of Ω and $a_1, ..., a_n \in \mathbb{R}_{\geq 0}$ then*
$$f = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i} \implies \int f d\mu = \sum a_i \mu(A_i)$$
3. *If $f \geq 0$ then*
$$\int f d\mu = \sup\{\int g d\mu : g \leq f, g \text{ is simple}\}$$

4. *Suppose we have $f = f^+ - f^-$ where $f^+, f^- \geq 0$. If $\int f^+ d\mu < \infty$ and $\int f^- d\mu < \infty$, i.e. $\int |f| d\mu < \infty$, then we say that f is integrable and define*

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

(We can actually allow one of the two to be infinite, as long as both are not, so we avoid the problem of $\infty - \infty$. However, we mostly focus on integrable functions.)

Proposition 3.2. *It is a proposition that the above constructions are well-defined and consistent with each other.*

Proposition 3.3. *Let f, g, f_n be nonnegative simple functions. Then*

1. $\int f d\mu \geq 0$
2. *For all $a \geq 0$, $\int (af) d\mu = a \int f d\mu$*
3. $\int (f+g) d\mu = \int f d\mu + \int g d\mu$
4. $f \geq g \implies \int f d\mu \geq \int g d\mu$
5. *If f_n is an increasing convergent sequence of functions then*

$$\lim_{n \to \infty} f_n \geq g \implies \lim_{n \to \infty} \int f_n d\mu \geq \int g d\mu$$

Proposition 3.4. *We have the following statements.*

1. *If $f_n \nearrow f$ where the f_n are nonnegative simple, then*

$$\int f d\mu = \lim_{n \to \infty} \int f_n d\mu$$

2. *In particular, we have*

$$\int f d\mu = \lim_{n \to \infty} \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mu(\mathbb{I}_{\{\frac{k}{2^n} \leq f < \frac{k+1}{2^n}\}}) + n\mu(\{f \geq n\})$$

3. *If $f \geq 0$ is measurable then $\int f d\mu \geq 0$*
4. *If $f \geq 0$ is measurable and $a \geq 0$ then $\int af d\mu = a \int f d\mu$*
5. *If $f, g \geq 0$ are measurable then $\int f + g d\mu = \int f d\mu + \int g d\mu$*
6. *If $f, g \geq 0$ are measurable and $f \leq g$ then $\int f d\mu \leq \int g d\mu$*

(The first point can be seen as an equivalent definition of the integral for nonnegative measurable functions because such a sequence always exists, as seen in the second point.)

Proposition 3.5. *Let f, g be integrable. Then for any $a \in \mathbb{R}$,*

$$\int (af) d\mu = a \int f d\mu$$

$$\int (f+g) d\mu = \int f d\mu + \int g d\mu$$

Definition 3.5. *For any $A \in \mathcal{F}$, we define*

$$\int_A f d\mu = \int f \mathbb{I}_A d\mu$$

Proposition 3.6. *We have the following properties:*

1. *If $\mu(A) = 0$ then for any measurable f, $\int_A f d\mu = 0$*
2. *If $f \leq g$ almost everywhere then $\int f d\mu \leq \int g d\mu$*
3. *If $f = g$ a.e. then $\int f d\mu = \int g d\mu$*

4 Convergence Theorems

Now, we consider the problem of exchanging limits and integration. We have the following major results:

Theorem 4.1 (**Monotone Convergence Theorem**). *Suppose we have a convergent sequence of measurable functions $f_n \nearrow f$ (pointwise) where $f_n, f \geq 0$. Then*

$$\lim_{n \to \infty} \int f_n d\mu = \int f d\mu$$

Lemma 4.2 (**Fatou's Lemma**). *If $f_n \geq 0$ then*

$$\int \liminf_n f_n d\mu \leq \liminf_n \int f_n d\mu$$

Theorem 4.3 (**Dominated Convergence Theorem**). *Suppose $f_n \to f$ and there exists an integrable function g such that $|f_n| \leq g$ for all $n \in \mathbb{N}$. Then*

$$\lim_{n \to \infty} \int f_n d\mu = \int f d\mu$$

Corollary 4.3.1. *If μ is finite, $f_n \to f$ and $|f_n| \leq M \in \mathbb{R}$, then*

$$\lim_{n \to \infty} \int f_n d\mu = \int f d\mu$$

5 Strong Law of Large Numbers

Now, we move to the case of probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. We begin with some definitions.

Remark: In this setting, we define $\mathbb{E}[X] = \int X d\mathbb{P}$, whenever X is integrable.

Definition 5.1 (**Independence** (Events))**.** *Given events $A_1, A_2, ..., A_n \in \mathcal{F}$, we say they are independent if for all $S \subseteq \{1, 2, ..., n\}$,*

$$\mathbb{P}(\bigcap_{i \in S} A_i) = \prod_{i \in S} \mathbb{P}(A_i)$$

Definition 5.2 (**Independence** (Set families))**.** *Consider $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_n \subseteq \mathcal{F}$. We say they are independent if for all choices of $A_i \in \mathcal{E}_i$, the events $A_1, A_2, ..., A_i$ are independent.*

Definition 5.3. *Given a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, we denote by $\sigma(X)$ the smallest σ-algebra such that $X : (\Omega, \sigma(X), \mathbb{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is measurable.*

Remark: We can see that $\sigma(X) = X^{-1}\mathcal{B}_{\mathbb{R}}$, and we know that $\mathcal{B}_{\mathbb{R}} = \sigma\{[a, \infty) : a \in \mathbb{Q}\}$. Thus, we have

$$\sigma(X) = \sigma(\{\{X \geq a\} : a \in \mathbb{Q}\})$$

For example, if $X = \mathbb{I}_A$, then $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$.

Definition 5.4 (**Independence** (Random variables))**.** *If $X_1, X_2, ..., X_n$ are random variables, we say they are independent if $\sigma(X_1), \sigma(X_2), ..., \sigma(X_n)$ are independent as set families.*

Proposition 5.1. *If X, Y are independent and X, Y, and XY are integrable, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

The following propositions were useful in the proofs of SLLN:

Lemma 5.1 (**Borel-Cantelli**)**.** *If A_n are events such that*

$$\sum_n \mathbb{P}(A_n) < \infty$$

then with probability 1, the A_n do not occur infinitely often. That is,

$$\mathbb{P}[\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n] = 0$$

Proposition 5.2 (**Markov Inequality**)**.** *If $X \geq 0$ is a random variable and $a > 0$ then*

$$\mathbb{P}(X \geq a) \leq \frac{E[X]}{a}$$

Proposition 5.3 (**Chebyshev Inequality**)**.** *If X is an integrable random variable with variance $\sigma^2 \neq 0$ and mean μ, then for any $a > 0$,*

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Theorem 5.2 (**SLLN** Version 1). *Suppose $X_1, X_2, ...$ are independent random variables such that $\mathbb{E}[X_i] = \mu \in \mathbb{R}$ for all $i \in \mathbb{N}$ and $\max_i \mathbb{E}[X_i^4] < \infty$. Then with probability 1,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

Theorem 5.3 (**SLLN** Version 2). *Suppose $X_1, X_2, ...$ are independent random variables and for each i, $\mathbb{E}[X_i] = 0$. Moreover, suppose that $\max_{i\in\mathbb{N}} \mathbb{E}[X_i^2] < \infty$. Then with probability 1,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

The following lemma was useful in the proof of SLLN Version 2:

Lemma 5.4 (**Maximal Inequality**). *Suppose $Z_1, Z_2, ..., Z_N$ are independent, let $\varepsilon_1, \varepsilon_2 > 0$, and define*

$$\beta^{-1} = \min_{1\le i\le N} \mathbb{P}(|Z_{i+1} + ... + Z_N| \le \varepsilon_2)$$

Then we have

$$\mathbb{P}(\max_{1\le i\le N} |Z_1 + ... + Z_i| > \varepsilon_1 + \varepsilon_2) \le \beta\mathbb{P}(|Z_1 + Z_2 + ... + Z_N| > \varepsilon_1)$$

Theorem 5.5 (**Kolmogorov's SLLN**). *Suppose $X_1, X_2, ...$ are independent random variables and for each i, $\mathbb{E}[X_i] = 0$. Moreover, suppose that*

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[X_i^2]}{i^2} < \infty$$

Then with probability 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

Theorem 5.6 (**SLLN** Version 3). *Suppose $X_1, X_2, ...$ are i.i.d random variables and $\mathbb{E}[X_i] = 0$. Then with probability 1,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\infty} X_i = 0$$

6 Weak Convergence

To define weak convergence, we move to a more particular setting. Let $(\mathcal{X}, d)$ be a metric space. Then it has a topology, and hence a Borel σ-algebra. We may thus define a measurable space $(\mathcal{X}, B(\mathcal{X}))$.

Definition 6.1 (Distribution). *Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, B(\mathcal{X}))$ be a random variable. Then X induces a measure P on $\mathcal{X}$ such that for any Borel set B,*

$$P(B) = \mathbb{P}(X^{-1}B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

We call P the pushforward measure, or the distribution of X, and we write $X \sim P$.

The above definition is why we may forget about the domains of random variables: they are fully characterized by their distribution. Now, for any measurable function $f : (\mathcal{X}, B(\mathcal{X})) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, we can write $Pf = \int f dP$, which is the integral of f w.r.t the distribution P. Now, we can define

Definition 6.2. *The class of bounded and Lipschitz functions on $\mathcal{X}$ is defined as*

$$BL(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \mid \sup_x |f(x)| < \infty, \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} < \infty\}$$

For a given $f \in BL(\mathcal{X})$, the value of the second supremum is called the Lipschitz constant and denoted $L \in \mathbb{R}$.

Definition 6.3 (Weak convergence). *Let $(P_n)_n$ be a sequence of distributions and let P be a distribution such that for all $f \in BL(\mathcal{X})$, $P_n f \to Pf$. Then we say that the P_n converge weakly to P and write $P_n \rightsquigarrow P$. Moreover, for a sequence of random variables $(X_n)_n$ and a distribution P, we can write $X_n \rightsquigarrow P$ if $P_n \rightsquigarrow P$ where P_n is the distribution of X_n. Alternatively, we can write $X_n \rightsquigarrow X$ if $P_n \rightsquigarrow P$ where P_n is the distribution of X_n and P is the distribution of X.*

Note: even if $X_n \to X$, the X_n may have different domains. We only require that they all have the same codomain, some metric space $\mathcal{X}$ with its Borel σ-algebra. (For our purposes, this will always be $\mathbb{R}$.) Now, we will give some equivalent conditions for weak convergence.

Definition 6.4 (Lower semi-continuous)**.** *We say a function $g : \mathcal{X} \to \mathbb{R}$ is lower semi-continuous (LSC) if for all $t \in \mathbb{R}$, the set $\{g > t\}$ (i.e. $\{x \in \mathcal{X} : g(x) > t\}$) is open.*

Ex: for indicator functions, $\mathbb{I}_B$ is LSC exactly for open sets B.

Definition 6.5 (Upper semi-continuous)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is called upper semi-continuous (USC) if $-f$ is LSC, i.e. for all $t \in \mathbb{R}$, $\{f \geq t\}$ is closed.*

Lemma 6.1. *Suppose g is LSC and $g \geq -M$ for some $M > 0$. Then there exists a sequence $l_k \in BL(\mathcal{X})$ such that $l_k \nearrow g$.*

With this lemma, we proved the following theorem:

Theorem 6.2. *Suppose $P_n \rightsquigarrow P$. Then*

- *For all LSC functions g such that $g \geq -M$, $\liminf_n P_n g \geq Pg$*
- *For all USC functions f such that $f \leq M$, $\limsup_n P_n f \leq Pf$*

Now, to find more equivalent conditions for weak convergence, we give new definitions.

Definition 6.6. *Given a function $f : \mathcal{X} \to \mathbb{R}$, we define*

$$f^\circ = \sup\{g : \mathcal{X} \to \mathbb{R} \mid g \leq f, g \text{ is LSC}\}$$

Ex: if $f = \mathbb{I}_B$ is an indicator function, then $(\mathbb{I}_B)^\circ = \mathbb{I}_{\text{int}(B)}$.

Definition 6.7. *Given a function $f : \mathcal{X} \to \mathbb{R}$, we define*

$$\overline{f} = \inf\{g : \mathcal{X} \to \mathbb{R} \mid g \geq f, g \text{ is USC}\}$$

Analogously, if we take $f = \mathbb{I}_B$ for some set B, then $\overline{f}$ is the indicator of $\overline{B}$, the closure. We note that for any function f, $f^\circ \leq f \leq \overline{f}$.

Proposition 6.1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is bounded above and below. Then for any x s.t. $f^\circ(x) = \overline{f}(x)$, f is continuous at x. Moreover, if f is continuous at x, then $f^\circ(x) = \overline{f}(x)$.*

These observations yield the following theorem:

Theorem 6.3. *If $P_n \rightsquigarrow P$, then for any bounded measurable f that is continuous almost everywhere (w.r.t the measure P), we have $P_n f \to Pf$.*

Corollary 6.3.1. *Suppose $P_n \rightsquigarrow P$. Then*

- *For any open G, $\liminf_n P_n(G) \geq P(G)$*
- *For any closed F, $\limsup_n P_n(F) \leq P(F)$*
- *For any Borel set B such that $P(\partial B) = 0$, $P_n(B) \to P(B)$*

These results can be combined into the following theorem:

Theorem 6.4. *Given a sequence of distributions P_n and a distribution P, the following are equivalent conditions:*

- *For any $l \in BL(\mathcal{X})$, $P_n l \to Pl$*
- *For any LSC f such that $f \geq -M$, $\liminf_n P_n f \geq Pf$*
- *For any USC f such that $f \leq M$, $\limsup_n P_n f \leq Pf$*
- *For any open set G, $\liminf_n P_n(G) \geq P(G)$*
- *For any closed set F, $\liminf_n P_n(F) \leq P(F)$*
- *For any Borel set B such that $P(\partial B) = 0$, $P_n(B) \to PB$*
- *If $-M \leq f \leq M$ and f is continuous almost everywhere (w.r.t P) then $P_n f \to Pf$*

Now, we give a brief discussion of product spaces. Given two measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, we have $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_i \in \Omega_i\}$. Then $\mathcal{F}_1 \times \mathcal{F}_2$ is the set of measurable rectangles, but it is not yet a σ-algebra. Thus, we instead define

$$\mathcal{F}_1 \times \mathcal{F}_2 := \sigma(\{A_1 \times A_2 : A_i \in \mathcal{F}_i\})$$

To define the measure on this new measurable space, we have the following theorem:

Theorem 6.5 (Fubini's Theorem). *Suppose we are in the above setting and μ_1, μ_2 are finite. Then there exists a unique measure μ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$, denoted by $\mu_1 \times \mu_2$, such that for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$,*

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$$

Moreover, for any f s.t. $\int |f| d\mu < \infty$, we have

$$\int f d\mu = \int \int f(\omega_1, \omega_2) d\mu_1(\omega_1)\mu_2(\omega_2) = \int \int f(\omega_1, \omega_2) d\mu_2(\omega_2)\mu_1(\omega_1)$$

We now give one final theorem about weak convergence. We define the set of smooth functions by

$$C^\infty(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \mid \forall k \in \mathbb{N}_{\geq 0}, \sup_{x\in\mathbb{R}} |f^{(k)}(x)| < \infty\}$$

(That is, f has all derivatives and all are bounded.) With this, we have the following theorem.

Theorem 6.6. *If P_n is a sequence of distributions and P is a distribution, the following conditions are equivalent:*

- *For all $f \in BL(\mathbb{R})$, $P_n f \to Pf$*
- *For all $f \in C^\infty(\mathbb{R})$, $P_n f \to Pf$*
- *For all $x \in \mathbb{R}$ such that $P(\{x\}) = 0$, $P_n(0, x] \to P(0, x]$.*

7 Central Limit Theorems

Having studied and given various conditions for weak convergence, we are prepared to formulate and prove multiple versions of the CLT.

Proposition 7.1 (Lindeberg swapping trick). *Let $f \in C^\infty(\mathbb{R})$, and consider random variables $\xi_1, \xi_2, \ldots, \xi_k, \eta_1, \eta_2, \ldots, \eta_k$ such that the η_i are Gaussian and for $1 \leq i \leq k$, $\mathbb{E}[\xi_i] = \mathbb{E}[\eta_i]$ and $\mathbb{E}[\xi_i^2] = \mathbb{E}[\eta_i^2]$. Moreover, suppose that all random variables have finite third moment. Then*

$$|\mathbb{E}f(\sum_{i=1}^k \xi_i) - \mathbb{E}f(\sum_{i=1}^k \eta_i)| \leq M \sum_{i=1}^k \mathbb{E}(|\xi_i|^3)$$

for some universal constant M.

In the proof of this statement, we used the following results:

Lemma 7.1 (Hölder's Inequality). *For measurable functions f, g and $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\int |fg| d\mu \leq (\int |f|^p d\mu)^{\frac{1}{p}} (\int |g|^q d\mu)^{\frac{1}{q}}$$

Lemma 7.2 (Jensen's Inequality). *If X is an integral random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

In general, if we take a sum of n i.i.d. random variables X_i, in order to make sure they are normalized, the distribution of X_1 may change as $n \to \infty$. Thus, we actually require two indices. We therefore introduce triangular arrays:

Definition 7.1 (Triangular array)**.** *A triangular array is a collection of random variables $\xi_{i,j}$ arranged into a table with some conditions. The i-th row contains the r.v. $\xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,k(i)}$, where $k(i) \to \infty$ as $i \to \infty$. Moreover, we assume that within a row, the $\xi_{i,n}$ are independent. Between rows, the r.v. need not have any relation, or even share a domain.*

We are now ready to state one version of the CLT.

Theorem 7.3 (**Central Limit Theorem** Version 1). *Given a triangular array of random variables $\xi_{i,j}$, suppose we have the following:*

- $\sum_{i=1}^{k(n)} \mathbb{E}[\xi_{n,i}] \to \mu$ *as* $n \to \infty$
- $\sum_{i=1}^{k(n)} \mathbb{E}[\xi_{n,i}^2] \to \sigma^2$ *as* $n \to \infty$
- $\sum_{i=1}^{k(n)} \mathbb{E}(|\xi_{n,i}|^3)] \to 0$ *as* $n \to \infty$

Then we may conclude that

$$\sum_{i=1}^{k(n)} \xi_{n,i} \xrightarrow{n\to\infty} N(\mu, \sigma^2)$$

Corollary 7.3.1. *Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. random variables with $\mathbb{E}[X_i] = 0$ and $Var[X_i] = 1$ for each $i \in \mathbb{N}$. Then as $n \to \infty$,*

$$\frac{X_1 + \ldots + X_n}{\sqrt{n}} \rightsquigarrow N(0,1)$$

Theorem 7.4 (**Lindeberg's CLT**). *Suppose we have a triangular array of random variables $X_{n,i}$ such that:*

- $\mathbb{E}[X_{n,i}] = 0$ *for each* n, i
- $\sum_{i=1}^{k(n)} \mathbb{E}[X_{n,i}^2] = 1$ *for each* n
- $\sum_{i=1}^{k(n)} \mathbb{E}[X_{n,i}^2 \mathbb{I}\{|X_{n,i}| > \varepsilon\}] \xrightarrow{n\to\infty} 0$ *for each* $\varepsilon > 0$.

Then

$$\sum_{i=1}^{k(n)} X_{n,i} \xrightarrow{n\to\infty} N(0,1)$$

In this proof, we used the following lemma:

Lemma 7.5. *Suppose we have a sequence of functions $H_n(\varepsilon)$ such that for each $\varepsilon > 0$, $H_n(\varepsilon) \xrightarrow{n\to\infty} 0$. Then there exists a sequence $\varepsilon_n \to 0$ such that $H_n(\varepsilon_n) \to 0$ as well.*

Next, we examined an entirely different way of proving CLT, beginning with the following lemma:

Lemma 7.6 (**Stein's Identity**). *Suppose $Z \sim N(0,1)$. Then for any absolutely continuous function f such that the following are well-defined, we have*

$$\mathbb{E}[f'(Z)] = \mathbb{E}[Zf(Z)]$$

Definition 7.2. *We say that a function f is absolutely continuous if it is differentiable and for all a, b we have*

$$f(b) - f(a) = \int_a^b f'(t)d\lambda(t)$$

We also have a converse to the previous lemma:

Lemma 7.7. *Suppose we have a random variable W and for any absolutely continuous function s.t. $\mathbb{E}[|f'(Z)|] < \infty$, we have*

$$\mathbb{E}[f'(W)] = \mathbb{E}[Wf(W)]$$

Then $W \sim N(0,1)$.

The main idea of the proof was to find a solution f to the differential equation

$$f'(w) - wf(w) = \mathbb{I}\{w \le t\} - \Phi(t)$$

for any given t, where $\Phi(t)$ is the CDF of the normal distribution. Then, we could take the expectation of both sides and apply the hypothesis to see that

$$0 = \mathbb{E}[f'(W) - Wf(W)] = \mathbb{P}(W \le t) - \Phi(t) \implies \mathbb{P}(W \le t) = \Phi(t)$$

and conclude that W has the same CDF as Z, and thus $W \sim N(0,1)$. With this in mind, we have the following proposition:

Proposition 7.2 (Stein's Differential Equation). *Consider the differential equation*

$$f'(w) - wf(w) = h(w) - Nh$$

for a given function h, where $Nh = \mathbb{E}[h(Z)]$ when $Z \sim N(0,1)$. This has a solution given by

$$f_h(w) = e^{\frac{w^2}{2}} \int_{-\infty}^{w} (h(x) - Nh) e^{-\frac{x^2}{2}} dx$$

(We used the special case of $h = \mathbb{I}\{x \le t\}$.) Now, we can get quantitative bounds on convergence in the CLT. First, we have the following result about the solution to Stein's differential equation:

Lemma 7.8. *If $f_h(x)$ is the solution given above, then*

$$\sup_{x \in \mathbb{R}} |f_h(x)|, \ \sup_{x \in \mathbb{R}} |f_h'(x)|, \ \sup_{x \in \mathbb{R}} |f_h''(x)| \le 2Lip(h)$$

where $Lip(h)$ is the Lipschitz constant of h, if it has one.

In proving this result, we used the following inequality:

Proposition 7.3 (Chernoff bound). *For any random variable X, we have that for any $\lambda > 0$,*

$$\mathbb{P}(X \ge a) \le \mathbb{E}[e^{\lambda X}] e^{-\lambda a}$$

In particular, this holds for the minimum such value over all $\lambda > 0$, so solving that optimization problem gives a better bound.

To describe the rate of convergence, we need a way to quantiy distance between two distributions. One such way is the Wasserstein distance:

Definition 7.3 (**Wasserstein distance**). *Given distributions W and Z, we define the Wasserstein distance by*

$$\mathcal{W}(W,Z) = \sup_{h:Lip(h)\leq 1} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]|$$

We can now return to our particular setting and make use of our lemma to prove the following bound:

Proposition 7.4.

$$\begin{aligned}\mathcal{W}(W,Z) &= \sup_{h:Lip(h)\leq 1} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]| \\ &= \sup_{h:Lip(h)\leq 1} |\mathbb{E}(f_h'(W)) - \mathbb{E}(Wf_h'(W))| \\ &\leq \sup_{||f||_\infty,\, ||f'||_\infty,\, ||f''||_\infty \leq 2} |\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)]|\end{aligned}$$

Next, we used the leave-one-out method to prove the following bound:

Proposition 7.5. *Let $X_1, X_2, \dots, X_n$ be independent random variables such that $\mathbb{E}[X_i] = 0$ and $Var[X_i] = 1$ for all i. Then by setting $W = \frac{1}{\sqrt{n}} \sum X_i$ and applying the previous results, we have*

$$\mathcal{W}(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}, N(0,1)) \leq \frac{3}{n\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}(|X_i|^3)$$

Now, we define another distance metric between distributions.

Definition 7.4 (**Kolmogorov-Smirnov distance**). *For distributions W and Z, we define*

$$KS(W,Z) = \sup_{t\in\mathbb{R}} |\mathbb{P}(W \leq t) - \mathbb{P}(Z \leq t)|$$

(In other words, the KS distance is the infinity-norm of the difference of the two CDFs.) The two distances have the following relationship:

Proposition 7.6. *If W is some distribution and $Z \sim N(0,1)$ then*

$$KS(W,Z) \leq 2\sqrt{\mathcal{W}(W,Z)}$$

Combining this result with that of Proposition 7.4, if we put the same hypotheses on the X_i, then

$$KS(\frac{X_1 + \ldots + X_n}{\sqrt{n}}, N(0,1)) \leq 2\sqrt{\frac{3}{n\sqrt{n}} \sum \mathbb{E}(|X_i|^3)}$$

We stated but did not prove the following improvement on this result:

Proposition 7.7 (Berry-Esseen). *Suppose we have i.i.d random variables $X_1, X_2, \dots, X_n$ such that $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$, and $\mathbb{E}[|X^3|] < \infty$. Then*

$$KS(\frac{X_1 + \ldots + X_n}{\sqrt{n}}, N(0,1)) \leq \frac{C\mathbb{E}[|X^3|]}{\sqrt{n}}$$

for some known (universal) constant C.

8 Coupling and Poisson Approximation

In covering examples of CLT, we saw that if the X_i are i.i.d. Bernoulli(p_n) random variables and $np(1-p) \to \infty$, then $\sum X_i$, once normalized, can be well-approximated by a normal distribution. Suppose instead that $np \to \lambda$ for some fixed λ. Then in fact $\sum X_i$ is better approximated by a Poisson distribution. To build out similar results for Poisson distributions, we begin by discussing couplings.

If random variables W and Z have the same domain, we can directly compare them, and perform operations. But what if we are just given distributions? How can we construct random variables according to those distributions that have the same domain? That is exactly the problem of finding a coupling:

Definition 8.1. *Given random variables X and Y, a coupling is a pair $(\hat{X}, \hat{Y})$ such that $\hat{X}$ and $\hat{Y}$ have the same domain, $\hat{X} \sim X$, and $\hat{Y} \sim Y$ (i.e. they have the same distribution).*

For example, suppose $X, Y, U \sim \text{Unif}([0,1])$. Then (U, U) is a coupling of X and Y, but so is $(U, 1-U)$. But intuitively, the coupling (U, U) better demonstrates that they have the same distribution, since the difference is zero everywhere. To formalize this concept, we give the following definition:

Definition 8.2. *Given random variables X and Y, the coupling $(\hat{X}, \hat{Y})$ is called maximal if it maximizes $\mathbb{P}(\hat{X} = \hat{Y})$ (over the space of all possible couplings).*

Theorem 8.1. *If X and Y are random variables taking values in $\mathbb{N}_{\geq 0}$, we can define $p_n = \mathbb{P}(X = n)$ and $q_n = \mathbb{P}(Y = n)$. The maximal coupling then satisfies*

$$\mathbb{P}(\hat{X} = \hat{Y}) = \sum_{n \in \mathbb{N}_{\geq 0}} \min(p_n, q_n)$$

Definition 8.3 (Total Variation Distance). *If P and Q are distributions, the total variation distance is defined by*

$$TV(P, Q) = \sup_{B \in \mathcal{B}_{\mathbb{R}}} |P(B) - Q(B)|$$

Lemma 8.2. *Let $p = \frac{dP}{d(P+Q)}$ and $q = \frac{dQ}{d(P+Q)}$ be the Radon-Nikodym derivatives. Then*

$$TV(P, Q) = \frac{1}{2} \int |p - q| d(P + Q) = 1 - \int \min(p, q) d(P + Q)$$

Theorem 8.3 (Le Cam's Theorem). *Suppose we have independent random variables $X_i \sim$ Bernoulli(p_i) for $1 \leq i \leq n$. We can define $Z \sim$ Poiss(λ) where $\lambda = \sum p_i$. Then*

$$TV(Z, \sum_{i=1}^{n} X_i) \leq \sum_{i=1}^{n} p_i^2$$

Example: Let $p_i = \frac{\lambda}{n}$ for a given constant λ. Then we have $\mathrm{TV}(Z, \sum X_i) \leq \frac{\lambda^2}{n} \to 0$ as $n \to \infty$. In this case, $\sum X_i$ is a binomial distribution with fixed expected value λ. Thus, we find that a binomial distribution with fixed expectation converges to a Poisson distribution as $n \to \infty$. (On the other hand, a normalized binomial distribution with fixed probability p converges to a normal distribution.)

However, we can attain a stronger bound.

Lemma 8.4 (Stein-Chen). *Suppose $Z \sim$ Poiss(λ). Then*

$$\mathbb{E}[Zf(Z)] = \lambda \mathbb{E}[f(Z+1)]$$

Moreover, if there is a random variable W such that the above holds for all bounded functions $f : \mathbb{N}_{\geq 0} \to \mathbb{R}$. Then $W \sim$ Poiss(λ).

The structure of the argument for the converse part is similar to Stein's method for CLT, and dealing with normal distributions. The more general formulation of the Stein-Chen difference equation, analogous to Stein's differential equation, is:

$$\lambda f(k+1) - kf(k) = \mathbb{I}_{\{k \in B\}} - \mathbb{P}(Z \in B)$$

for some subset $B \subseteq \mathbb{N}_{\geq 0}$. Proving the converse above involved finding a solution when $B = \{t\}$ for arbitrary $t \in \mathbb{N}_{\geq 0}$. The more general solution is given by

$$f_B(k+1) = \frac{k!}{\lambda^{k+1}} \sum_{j=0}^{k} \frac{\lambda^j}{j!} [\mathbb{I}\{j \in B\} - \mathbb{P}(Z \in B)]$$

Lemma 8.5. *If f_B is the solution to the above difference equation, then*

$$f_B(k) - f_B(l) \leq \min(\frac{1}{\lambda}, 1)|k - l|$$

As a corollary,

Corollary 8.5.1. *If $Z \sim$ Poiss(λ), then*

$$TV(W, Z) \leq \sup_{f : |f(k) - f(l)| \leq 1} |\mathbb{E}[Wf(W)] - \lambda \mathbb{E}[f(W+1)]|$$

We can use this to improve on the earlier result of Le Cam:

Theorem 8.6. *If $X_i \sim$ Bernoulli(p_i) are independent random variables for $1 \leq i \leq n$, we can define $\lambda = \sum p_i$ and $Z \sim$ Poiss(λ). Then*

$$TV(Z, \sum X_i) \leq \min(1, \frac{1}{\lambda}) \sum_{i=1}^{n} p_i^2$$

This concludes the section on Poisson approximation.

9 Measure Decomposition, Radon-Nikodym Derivative

The general question of the Radon-Nikodym derivative is as follows: given measures φ and μ on $(\Omega, \mathcal{F})$, is there a measurable f such that

$$\varphi(A) = \int_A f d\mu$$

for all $A \in \mathcal{F}$? If so, f is denoted by $\frac{d\varphi}{d\mu}$. This is not always possible, as we can take $\mu \sim \text{Unif}([0,1])$ and $\varphi \sim \text{Unif}([0,2])$. Then $\int_{[1,2]} f d\mu = 0$ for all f, but $\varphi([1,2]) = 0.5$.

To answer this question, we begin with the following definition:

Definition 9.1. *A function φ is called a signed measure on $(\Omega, \mathcal{F})$ if $\varphi(\emptyset) = 0$ and for any countable collection of disjoint sets $A_n \in \mathcal{F}$,*

$$\varphi(\bigcup_n A_n) = \sum_n \varphi(A_n)$$

Note that unlike measures, φ may take on positive or negative values. As we prove in homework, φ may take on the values $+\infty$ or $-\infty$, but never both.

If for any $A \in \mathcal{F}$, $|\varphi(A)| < \infty$, then φ is called a finite signed measure. As we see in the following theorem, signed measures can be reduced to the sum of two measures:

Theorem 9.1 (**Hahn Decomposition Theorem**). *Let φ be a signed measure on $(\Omega, \mathcal{F})$. Then there exists a partition $\Omega = \Omega_+ \sqcup \Omega_-$ such that for all $A \in \mathcal{F}$, $\varphi(A \cap \Omega_+) \geq 0$ and $\varphi(A \cap \Omega_-) \leq 0$. Moreover, the decomposition is unique in the following sense: suppose we have two such decompositions*

$$\Omega_{1,+} \sqcup \Omega_{1,-} = \Omega = \Omega_{2,+} \sqcup \Omega_{2,-}$$

Then for any $A \in \mathcal{F}$ such that $A \subseteq \Omega_{1,+} \Delta \Omega_{2,+}$, $\varphi(A) = 0$. (And the same holds for the other pair.)

Now, for any A, we can write $\varphi(A) = \varphi^+(A) - \varphi^-(A)$ where $\varphi^+(A) = \varphi(A \cap \Omega_+)$ and $\varphi^-(A) = \varphi(A \cap \Omega_-)$. In particular, φ^+ and φ^- are measures. This is called the Jordan decomposition of φ. Now, we move to the Radon-Nikodym derivative.

Definition 9.2 (**Absolute continuity**). *Given a signed meausre φ and a measure μ, suppose that for any $A \in \mathcal{F}$,*

$$\mu(A) = 0 \implies \varphi(A) = 0$$

Then we say that φ is absolutely continuous w.r.t μ and write $\varphi << \mu$.

Theorem 9.2 (**Radon-Nikodym Theorem**). *Suppose φ is a signed measure and μ is a σ-finite measure such that $\varphi << \mu$. Then there exists an (almost everywhere) unique measurable function f such that for all $A \in \mathcal{F}$,*

$$\varphi(A) = \int_A f d\mu$$

Then f is called the Radon-Nikodym derivative and is denoted by $\frac{d\varphi}{d\mu}$.

In lectures, we proved the special case when both φ and μ are finite. We utilized an additional lemma that I exclude here because it was just an intermediate step in that proof and doesn't have clear applications to problem-solving.

10 Conditional Expectation

Definition 10.1 (**Conditional Expectation**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra, and let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ be an integrable random variable. Then we define the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ to be the unique function that is measurable w.r.t $\mathcal{G}$ and such that for all $A \in \mathcal{G}$,*

$$\int_A \mathbb{E}[X|\mathcal{G}] d\mathbb{P} = \int_A X d\mathbb{P}$$

For one construction, we can define a signed measure on $(\Omega, \mathcal{G}, \mathbb{P})$ by

$$\varphi(A) = \int_A X d\mathbb{P}$$

for any $A \in \mathcal{G}$. Then $\varphi << \mathbb{P}$, so there exists a Radon-Nikodym derivative $\frac{d\varphi}{d\mathbb{P}} : (\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}_R)$ that is measurable and such that for all $A \in \mathcal{G}$,

$$\varphi(A) = \int_A X d\mathbb{P} = \int_A \frac{d\varphi}{d\mathbb{P}} d\mathbb{P}$$

Then we can see that $\frac{d\varphi}{d\mathbb{P}}$ is by definition the conditional expectation $\mathbb{E}[X|\mathcal{G}]$. According to the Radon-Nikodym Theorem, this is unique except for $\mathbb{P}$-null sets in $\mathcal{G}$. Before giving this construction, we examined some desired properties of conditional expectation.

First, if $X = \mathbb{I}_A$ and $Y = \mathbb{I}_B$ for $A, B \in \mathcal{F}$, then the conditional expectation, then the conditional expectation is binary, and takes on values according to the conditional probabilities:

$$\mathbb{E}[X|Y] = \mathbb{P}(A|B)\mathbb{I}\{Y = 1\} + \mathbb{P}(A|B^c)\mathbb{I}\{Y = 0\} = \mathbb{P}(A|B)\mathbb{I}_B + \mathbb{P}(A|B^c)\mathbb{I}_{B^c}$$

In fact, the above formula is correct if Y is replaced by the more generic $a\mathbb{I}_B + b\mathbb{I}_{B^c}$ for any $a \neq b$. In other words, the conditional expectation $\mathbb{E}[X|Y]$ depends only on $\sigma(Y)$, and not the values of Y itself. In this way, we capture the notion of the information provided by Y, and it is why we define the conditional expectation w.r.t a σ-algebra.

Proposition 10.1 (**Tower Law**). *If $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a random variable, and Z is a random variable that is measurable w.r.t $\mathcal{G}$, then*

$$\mathbb{E}[XZ] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]Z]$$

Equivalently, for any $A \in \mathcal{G}$,

$$\mathbb{E}[X\mathbb{I}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{I}_A]$$

Now, we prove some other properties of conditional expectation. I use 'a.s.' to mean almost surely w.r.t $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. up to a $\mathbb{P}$-null set in $\mathcal{F}$. Analogously, I use 'a.s..' to mean almost surely w.r.t $(\Omega, \mathcal{G}, \mathbb{P})$. (Note: measurability w.r.t $\mathcal{G}$ implies measurability w.r.t $\mathcal{F}$, while a.s. implies a.s.., because all null sets in $\mathcal{G}$ are null sets in $\mathcal{F}$.)

Example: If we take $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$. This represents the case of no information, and in fact X and $\mathcal{G}$ are independent.

Example: If we have $\mathcal{G} = \{\emptyset, \Omega, B, B^c\}$ and $X = \mathbb{I}_D$ then

$$\mathbb{E}[\mathbb{I}_D|\mathcal{G}] = \mathbb{P}(D|B)\mathbb{I}_B + \mathbb{P}(D|B^c)\mathbb{I}_{B^c}, a.s..$$

Example: If $\mathcal{G} = \mathcal{F}$, then X is already $\mathcal{G}$-measurable, so $\mathbb{E}[X|\mathcal{G}] = X$ a.s. This represents the case of total knowledge, and in general, larger σ-algebras $\mathcal{G}$ represent more information.

We formally defined conditional expectation w.r.t a sub-σ-algebra $\mathcal{G}$, but are often interested in $\mathbb{E}[X|Y]$ for random variables X and Y. In this case, we simply define

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

Lemma 10.1 (Doob-Dynkin Lemma). *If $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a random variables, then we have $\sigma(Y) \subseteq \mathcal{F}$. Suppose we also have $Z : (\Omega, \sigma(Y)) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, so that $\sigma(Z) \subseteq \sigma(Y)$. Then there exists a measurable function $f : (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ such that $Z = f(Y)$.*

Corollary 10.1.1. *As an immediate corollary, given random variables X and Y, there exists a measurable function $f : (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ such that*

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

Proposition 10.2. *For any sub-σ-algebra $\mathcal{G}$, random variable Z that is measurable w.r.t $\mathcal{G}$, and arbitrary integrable random variable X, we have*

$$\mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - X)Z] = 0$$

If we think of $\mathbb{E}[XY]$ as an inner product $\langle X, Y\rangle$ on the space of random variables, then we note that all Z are orthogonal to $\mathbb{E}[X|\mathcal{G}] - X$. We may thus think of $\mathbb{E}[X|\mathcal{G}]$ as the projection of X onto the subspace spanned by Z with $\sigma(Z) \subseteq \mathcal{G}$. There was more material on this idea in lecture, but it doesn't seem relevant to problem-solving, so I omit it from these notes.

Now, we have some more properties of conditional expectation:

Proposition 10.3. *The following are true:*

- $\sigma(X) \subseteq \mathcal{G} \implies \mathbb{E}[X|\mathcal{G}] = X$ *a.s..*
- *If $X \perp \mathcal{G}$ then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ a.s..*
- *If $\mathcal{G}_1 \subseteq \mathcal{G}_0$ then the following holds almost surely w.r.t. $\mathcal{G}_1$:*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_0]|\mathcal{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_0] = \mathbb{E}[X|\mathcal{G}_1]$$

- *If $X \leq Y$ a.s. then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$ a.s..*
- $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$

The above generalize some of the basic properties of integration (unconditional expectation) that we covered earlier in the course. Finally, we have analogous theorems about interchanging limits and integration.

Theorem 10.2. *Let $\mathcal{G} \subseteq \mathcal{F}$ be any sub-σ-algebra.*

- *If $0 \leq X_n \nearrow X$ a.s. and the X_n are integrable, then*
$$\mathbb{E}[X|\mathcal{G}] = \lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}], a.s..$$
- *If X_n is any sequence of integrable random variables then*
$$\mathbb{E}[\liminf_n X_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}], a.s..$$
- *Suppose $|X_n| \leq Y$ a.s. for some integrable Y. Then if $X_n \to X$ a.s., we have*
$$\mathbb{E}[X|\mathcal{G}] = \lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]$$

Finally, we have one last theorem:

Theorem 10.3. *Suppose $\sigma(Y) \subseteq \mathcal{G}$. Then for any integrable random variable X,*
$$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}], a.s..$$

11 Exercises

The following are a few homework problems that may be useful for solving problems.

Proposition 11.1. *Given a sequence of sets $A_n \in \mathcal{F}$,*
$$\mu(\liminf A_n) \leq \liminf \mu(A_n)$$
$$\mu(\bigcup_{n=1}^{\infty} A_n) < \infty \implies \mu(\limsup A_n) \geq \limsup \mu(A_n)$$

Proposition 11.2. *If $f \geq 0$ a.e. and $\int f d\mu = 0$ then $f = 0$ a.e.*

Proposition 11.3. *Suppose f and g are integrable functions such that for all $A \in \mathcal{F}$, $\int_A f d\mu \geq \int_A g d\mu$. Then $f \geq g$ almost everywhere.*

Proposition 11.4. *If X is a random variable then*
$$\mathbb{E}[|X|] < \infty \iff \sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) < \infty$$

Proposition 11.5. *Consider random variables on $\mathbb{R}$. We have the following:*

- $X_n \xrightarrow{\mathbb{P}} X$ *implies* $X_n \rightsquigarrow X$

- *If* $X_n \rightsquigarrow P$ *and* $Y_n \xrightarrow{\mathbb{P}} 0$ *then* $X_n + Y_n \rightsquigarrow P$
- *If* $X_n \rightsquigarrow P$ *and* $Y_n \xrightarrow{\mathbb{P}} 1$ *then* $X_n Y_n \rightsquigarrow P$ *and* $X_n / Y_n \rightsquigarrow P$
- *If* $X_n \rightsquigarrow X$ *and* $X = a \in \mathbb{R}$ *a.s. then* $X_n \xrightarrow{\mathbb{P}} a$

Proposition 11.6. *Suppose* μ *and* ν *are* σ*-finite measures such that* $\nu << \mu$*. For any* ν*-integrable function* f*,*

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