

[ORF] (526)

PROBABILITY THEORY

DR. MIKLOS RACZ • (FALL) 2018 • PRINCETON UNIVERSITY

Last Revision: October 25, 2018

Table of Contents

1	9/13/2018	1
1.1	Exercise	1
2	9/18/2018	1
2.1	Problem from last week	1
2.2	Probability and Measure Theory	1
2.3	Measure Theory	2
2.4	Intersections and Unions of σ -Algebras	2
2.5	Topological Spaces, Borel σ -Algebras, Borel Sets	2
2.6	Measurable Spaces	2
3	9/20/2018	3
3.1	Probability Spaces (review)	3
3.2	Measures	3
3.3	Properties of Measures	3
3.4	Specification of Measures	4
3.5	Completeness and Negligible Sets	4
3.6	Functions	4
4	9/25/2018	5
4.1	Measurable Functions	5
4.2	Approximation of Measurable Functions	6
4.3	Integration	6
5	9/27/2018	6
5.1	The Compactification of \mathbb{R}	6
5.2	Last Time	7
5.3	Continuing	7
5.4	Lebesgue Integration	7
5.5	Integration Over a Set	7
5.6	Monotone Convergence Theorem	8
5.7	The Insensitivity of the Integral W.R.T Negligible Sets	8
5.8	Fatou's Lemma	8
5.9	Dominated Convergence Theorem	8

5.10	A note on these theorems...	8
5.11	Characterization of the Integral	8
5.12	Product Spaces	9
5.13	Absolute Continuity of Measures	9
6	10/2/2018	9
6.1	Products of Measure Spaces	9
6.2	Probability	10
6.3	Expected Value	10
6.4	Weak Law of Large Numbers	10
6.5	Markov's Inequality	10
6.6	Chebyshev's Inequality	10
6.7	General Markov	11
6.8	Chernoff Bound	11
6.9	Almost Sure Convergence	11
6.10	Strong Law of Large Numbers	11
7	10/4/18	11
7.1	Strong Law of Large Number	11
7.2	Borel-Cantelli Lemmas (Sufficient conditions for Almost Everywhere)	12
7.3	Proof of SLLN	12
8	10/16/2018	13
8.1	Last Time	13
8.2	Characteristic Functions	13
8.3	Characteristic Functions and Moments	14
8.4	Examples of Characteristic Functions	14
8.5	Inversions	15
9	10/18/2018	15
9.1	Last Time	15
9.2	Levy's Continuity Theorem	15
9.3	Central Limit Theorem Proof	15
9.4	Some Comments on Weak Convergence	16
9.5	General Strategy for Proving Weak Convergence (Prohorov)	16
9.6	Stochastic Processes	17
9.7	Markov Processes and Markov Chains	17
9.8	Properties of Transition Probabilities Matrix	18
10	10/25/2018	18
10.1	Markov Chains	18
10.2	Graphs	20

Abstract

These notes are intended as a resource for myself; past, present, or future students of this course, and anyone interested in the material. The goal is to provide an end-to-end resource that covers all material discussed in the course displayed in an organized manner. If you spot any errors or would like to contribute, please contact me directly.

1 9/13/2018

LLN and CLT.

For coin flip:

Weak Law of Large Numbers: $\forall \epsilon \lim_{n \rightarrow \infty} \mathbb{P}(|\frac{S_n}{n} - \frac{1}{2}| \geq \epsilon) = 0$

Strong LLN: $\mathbb{P}(\lim_{n \rightarrow \infty} \frac{S_n}{n} = \frac{1}{2}) = 1$

General Strong LLN: If $\{x_i\}_{i=1}^{\infty}$ are I.I.D. random variables such that $\mathbb{E}[|x_i|] < \infty$ then if $S_n = x_1 + \dots + x_n$, $\lim_{n \rightarrow \infty} \frac{S_n}{n} = \mathbb{E}[x_i] = \mu$.

1.1 Exercise

Come up with some $\{x_i\}_{i=1}^{\infty}$ such that (A) the law of large numbers does not hold and (B) the law of large numbers does hold but the central limit theorem does not.

Problem: Throw a (fair) die until you get a 6. What is the expected number of throws (including the one that gives us a 6) conditioned on the event that all throws give even numbers?

2 9/18/2018

2.1 Problem from last week

Problem: Throw a (fair) die until you get a 6. What is the expected number of throws (including the one that gives us a 6) conditioned on the event that all throws give even numbers?

P. Cuff: Let T be the first time when the throw is not a 2 or a 4.

$T \sim \text{Geo}(\frac{2}{3})$, $\mathbb{E}[T] = \frac{3}{2}$. However, T is independent of what the die roll actually is, so our solution is $\mathbb{E}[T | X_T = 6] = \mathbb{E}[T] = \frac{3}{2}$. This example demonstrates the importance of being careful when dealing with conditions.

2.2 Probability and Measure Theory

In 1933, Kolmogorov laid down the mathematical foundations of probability theory using measure theory. Measure theory allows us to unify the theory dealing with discrete and continuous probability cases, deals with mixtures of both discrete and continuous cases, and deals with cases that are neither discrete nor continuous. Measure theory allows us to work with much more exotic probability problems. Ex: Stochastic processes in path space.

2.3 Measure Theory

Def: A collection C of subsets of E (the universe) is called an algebra if:

- 1: $\emptyset \in C$
- 2: $A \in C \implies A^c \in C$
- 3: $A, B \in C \implies A \cup B \in C$.

C is a σ -algebra if, in addition to (1) and (2),

- 3': $A_1, A_2, \dots \in C \implies \bigcup_{i=1}^{\infty} A_i \in C$.

Examples:

- 1: $\mathcal{E} = \{\emptyset, E\}$ trivial σ -algebra
- 2: $\mathcal{E} = 2^E$ discrete σ -algebra

2.4 Intersections and Unions of σ -Algebras

Any (countable or uncountable) intersection of σ -algebras is a σ -algebra.

The union of two σ -algebras is not necessarily a σ -algebra.

Def: (Generated σ -algebra) Let \mathcal{C} be a collection of subsets of E . Take all σ -algebras that contain \mathcal{C} . Take their intersection. This σ -algebra is called the σ -algebra generated by \mathcal{C} , and is denoted by $\sigma(\mathcal{C})$.

2.5 Topological Spaces, Borel σ -Algebras, Borel Sets

If E is a topological space, and \mathcal{C} is the collection of all open sets of E , then $\sigma(\mathcal{C})$ is called the Borel σ -algebra. Its elements are called Borel sets. The Borel σ -algebra is denoted by \mathcal{B}_E or $\mathcal{B}(E)$.

2.6 Measurable Spaces

Def: A pair (E, \mathcal{E}) is a measurable space if \mathcal{E} is a σ -algebra on E . The sets in \mathcal{E} are called measurable sets.

Def: Let (E, \mathcal{E}) , (F, \mathcal{F}) are two measurable spaces. If $A \subset E$ and $B \subset F$ are measurable sets, then $A \times B$ is called a measurable rectangle.

Def: The product $(E \times F, \mathcal{E} \otimes \mathcal{F})$ where $\mathcal{E} \otimes \mathcal{F} = \sigma(\{A \times B \mid A \in \mathcal{E}, B \in \mathcal{F}\})$, is a measurable space.

Def: $\mu : \mathcal{E} \rightarrow \mathbb{R}^+$ is a measure on (E, \mathcal{E}) if

- 1: $\mu(\emptyset) = 0$
- 2: If $A_1, A_2, \dots \in \mathcal{E}$ are pairwise disjoint, then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

(2) is called countable additivity or σ -additivity.

Def: A probability measure is a measure μ such that $\mu(E) = 1$.

Def: A probability space is a triple (E, \mathcal{E}, μ) such that (E, \mathcal{E}) is a measurable space, and μ is a probability measure.

Probability spaces are often denoted $(\Omega, \mathcal{F}, \mathbb{P})$.

3 9/20/2018

3.1 Probability Spaces (review)

A probability space, written $(\Omega, \mathcal{F}, \mathbb{P})$, is composed of a sample space (Ω) , the events (\mathcal{F}) , a σ -algebra on (Ω) , and a probability measure (\mathbb{P}) .

3.2 Measures

Suppose (E, \mathcal{E}) is a measurable space.

Def: We say that $\mu : \mathbf{E} \rightarrow \overline{\mathbb{R}_+}$ ($\overline{\mathbb{R}_+}$ is the reals plus the point at positive infinity) such that

(1): $\mu(\emptyset) = 0$

(2): If $A_1, A_2, \dots \in \mathcal{E}$ are pairwise disjoint then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Examples:

(1): The Dirac Measure. $x \in E$, $\delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$

(2): The Counting Measure. $D \in E$, $\mu(A) = \#$ of points in $A \cap D$. If D is countable, then $\mu(A) = \sum_{x \in D} \delta_x(A)$.

(3): Discrete Measure. $D \subset E$ countable, $m(x)$ is some real value for every $x \in D$. $\mu(A) = \sum_{x \in D} m(x) \delta_x(A)$.

(4): The Uniform Measure on $\{1, 2, \dots, n\}$. The discrete measure with $m(x) = \frac{1}{n}$.

(5): The Lebesgue Measure. $Leb(A) = \text{length of } A$ where A is an interval.

3.3 Properties of Measures

Let (E, \mathcal{E}, μ) be a measure space.

(1): Finite Additivity. $A \cap B = \emptyset \implies \mu(A \cup B) = \mu(A) + \mu(B)$.

(2): Monotonicity. If $A \subseteq B$ then $\mu(A) \leq \mu(B)$. Note that this is clear because $B = A \cup (B \setminus A)$.

(3): Sequential Continuity. If $A_n \subset A$ and A_n converges to A as $n \rightarrow \infty$ then $\mu(A_n)$ converges to $\mu(A)$ from below.

(4): Boole's Inequality / Union Bound. $A_1, A_2, \dots \in \mathcal{E}$, $\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

We can prove this by creating a sequence of disjoint subsets of $\bigcup_{i=1}^{\infty} A_i$, then using sequential continuity.

(5): If $c > 0$, then $c\mu$ is also a measure, $(c\mu)(A) = c \cdot \mu(A)$.

(6): If μ_1, μ_2 are measures then $\mu_1 + \mu_2$ is a measure.

Def: If $\mu(E) < \infty$ then μ is called a finite measure.

Def: We say that a measure μ is σ -finite if there exists a measurable countable partition $\{E_n\}$ of E such that $\mu(E_n) < \infty$ for all n . Ex: Leb is σ -finite.

3.4 Specification of Measures

Thm: Let (E, \mathcal{E}) be a measurable space. Let μ and ν be two measures on (E, \mathcal{E}) with $\mu(E) = \nu(E) < \infty$. If μ and ν agree on a collection of subsets that is closed under intersections, that generate \mathcal{E} , then $\mu = \nu$.

Cor: If μ and ν are two probability measures on \mathbb{R} with the same cumulative distribution functions, then $\mu = \nu$.

Def: The cumulative distribution at a point x is $\mu([-\infty, x])$.

Assume that $\{x\} \in \mathcal{E}$ if $x \in E$. This is true of all standard measurable spaces.

Def: x is an atom of μ if $\mu(\{x\}) > 0$.

Def: μ is purely atomic if $\exists D \subset E$ such that $\forall x \in D, \mu(\{x\}) > 0$ and $\mu(E \setminus D) = 0$.

Def: μ is diffuse if it has no atoms. Ex: Leb .

Lemma: If μ is a σ -finite measure on (E, \mathcal{E}) then we can write $\mu = \lambda + \nu$ where λ is diffuse and ν is purely atomic.

3.5 Completeness and Negligible Sets

Def: (Negligible Set)

A measurable set A is negligible if $\mu(A) = 0$. An arbitrary subset of E is negligible if it is contained in a measurable set that is negligible.

Def: A measure space is complete if every negligible set is measurable.

Lemma: To make a measure space complete, take $\bar{\mathcal{E}} = \sigma(\mathcal{E} \cup \mathcal{N})$, where \mathcal{N} is the collection of negligible sets. $\forall A \subset \bar{\mathcal{E}}, A = B \cup N$ with $B \in \mathcal{E}, N \in \mathcal{N}$. Define $\bar{\mu}(A) = \mu(B)$. This is called the completion of the measure space, $(E, \bar{\mathcal{E}}, \bar{\mu})$. In the case of $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, Leb)$, the elements of $\bar{\mathcal{B}}_{\mathbb{R}}$ are called Lebesgue-measurable.

3.6 Functions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space. $\Omega = \{1, 2, 3, 4, 5, 6\}$, $X = \text{outcome mod } 5$. $X(\omega) = \omega \bmod 5$, $X : \{1, 2, 3, 4, 5, 6\} \rightarrow \{0, 1, 2, 3, 4\}$. $\mathbb{P}(X = 1) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = 1\})$.

... Basic set theory stuff.

Def: (Measurable Functions)

(E, \mathcal{E}) and (F, \mathcal{F}) are two measure spaces. $f : E \rightarrow F$ is measurable relative to \mathcal{E} and \mathcal{F} if $f^{-1}(A) \in \mathcal{E}$ for every $A \in \mathcal{F}$.

Theorem: (E, \mathcal{E}) , (F, \mathcal{F}) measurable spaces. $f : E \rightarrow F$ is measurable relative to \mathcal{E} and \mathcal{F} if and only if there exists a collection \mathcal{F}_0 of subsets of F such that $f^{-1}(B) \in \mathcal{E} \forall B \in \mathcal{F}_0$, and \mathcal{F}_0 generates \mathcal{F} .

Proof: Left as an exercise.

Theorem: Composition. Let (E, \mathcal{E}) , (F, \mathcal{F}) , (G, \mathcal{G}) be measure spaces. $f : E \rightarrow F$, $g : F \rightarrow G$. If f and g are measurable, then $g \circ f$ is measurable.

4 9/25/2018

4.1 Measurable Functions

Let (E, \mathcal{E}, μ) and (F, \mathcal{F}) be measure spaces. Let $f : E \rightarrow F$.

Def: f is measurable relative to \mathcal{E} and \mathcal{F} if $f^{-1}(B) \in \mathcal{E} \forall B \in \mathcal{F}$.

Generally, we will focus on measurable functions $f : E \rightarrow \mathbb{R}$ (Real Valued function), $f : E \rightarrow \overline{\mathbb{R}} = [-\infty, \infty]$ (Numerical function), or similar.

Def: $f : E \rightarrow \mathbb{R}$ is \mathcal{E} -measurable if it is measurable relative to \mathcal{E} and $\mathcal{B}_{\mathbb{R}}$.

Def: If E is a topological space and \mathcal{E} is the Borel σ -algebra, then we simply say that f is a Borel function.

Lemma: $f : E \rightarrow \mathbb{R}$ is \mathcal{E} -measurable, if and only if $f^{-1}((-\infty, r]) \in \mathcal{E}$ for all $r \in \mathbb{R}$.

Pf: From HW1: $\sigma(\{(-\infty, r] \mid r \in \mathbb{R}\}) = \mathcal{B}(\mathbb{R})$. Then it follows from claim stated last time wrt the inverse of a generating set.

Def: $f^+ := \max\{f, 0\}$, $f^- := -\min\{f, 0\}$. Note that $f = f^+ - f^-$.

Lemma: f is \mathcal{E} -measurable if and only if f^+ and f^- are \mathcal{E} -measurable. The proof is left as an exercise.

Def: Indicator Functions.

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Check: $\mathbb{1}_A$ is \mathcal{E} -measurable if and only if $A \in \mathcal{E}$.

Def: A function is simple if $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$, $a_i \in \mathbb{R}$. Where A_1, A_2, \dots, A_n are \mathcal{E} -measurable.

Def: Canonical form: $f = \sum_{j=1}^m b_j \mathbb{1}_{B_j}$ where $\{B_j\}$ is a partition of \mathcal{E} .

Conversely, if a function is \mathcal{E} -measurable and takes only finitely many real values, then it is a simple function.

If f and g are simple, then so are $f + g$, $f - g$, fg , f/g , $\max\{f, g\}$, $\min\{f, g\}$.

Theorem: The class of measurable functions is closed under limits.

Let $\{f_n\}$ be a sequence of \mathcal{E} -measurable functions then $\inf f_n$, $\sup f_n$, $\liminf f_n$, and $\limsup f_n$, defined pointwise, are \mathcal{E} -measurable.

Pf: For $\sup f_n = f$, we want to show that $f^{-1}(-\infty, r] \in \mathcal{E}$. Since intersections can be rewritten as unions, and $f(x) \leq r \iff f_n(x) \leq r \forall n$, we have

$$f^{-1}(-\infty, r] = \bigcap_{n=1}^{\infty} f_n^{-1}(-\infty, r]$$

But we know that $f_n^{-1}(-\infty, r] \in \mathcal{E}$ and since this is a countable intersection, $f^{-1}(-\infty, r] \in \mathcal{E}$.

4.2 Approximation of Measurable Functions

let $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. We can approximate f by a sequence of simple functions by subdividing \mathbb{R}_+ into a partition with partitions of length $\frac{1}{2^n}$. If $x \in A$, $f_n(x)$ is the lower bound of $f(A)$. (???)

Alternatively, let $d_n(x) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{[\frac{k-1}{2^n}, \frac{k}{2^n})} + n \mathbb{1}_{[n, \infty)}$. Then $f_n = d_n \circ f$.

Thm: A function f is \mathcal{E} -measurable if and only if it is the increasing limit of simple functions.

Note: Lookup monotone classes of functions.

4.3 Integration

Suppose (E, \mathcal{E}, μ) is a measure space. Define $f : E \rightarrow \mathbb{R}$. We want to find $\int f d\mu$. That is, the integral of f relative to the measure μ . We denote this $\mu f = \mu(f) = \int f d\mu = \int \mu(dx) f(x) = \int_E \mu(dx) f(x)$.

How we will do this is we will first define integrals over measure spaces for simple functions, then extend this definition by taking limits.

Def: (Integration)

If f is a simple function, $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$, where $\{A_i\}$ is a partition of E we define the integral as

$$\int f d\mu = \sum_{i=1}^n a_i \mu(A_i)$$

Now, suppose that f is a measurable positive function, and let $f_n = d_n \circ f$. Then $\int f d\mu = \lim_{n \rightarrow \infty} \int f_n d\mu$.

Finally, if $f = f^+ - f^-$, then $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$, provided that at least one of the two integrals on the right are finite. Otherwise, $\int f d\mu$ is undefined.

5 9/27/2018

5.1 The Compactification of \mathbb{R}

$\mathcal{B}(\overline{\mathbb{R}}) = \sigma(\mathcal{B}(\mathbb{R}) \cup \{+\infty\} \cup \{-\infty\})$.

$\overline{\mathbb{R}}$ has the topology where $A \subset \overline{\mathbb{R}}$ is open if $A \setminus \{-\infty, +\infty\}$ is open in \mathbb{R} .

5.2 Last Time

For (E, \mathcal{E}, μ) , $f : E \rightarrow \overline{\mathbb{R}}$, we defined $\mu(f) = \int f d\mu$.

5.3 Continuing

Def: f is integrable if $\int f d\mu$ exists and is finite.

Note: f is integrable $\iff \int |f| d\mu < \infty$. Notice $|f| = f^+ + f^-$.

Exercise: Every integrable function is real-valued almost everywhere (a.e.).

Def: A statement holds almost everywhere (for almost every $x \in E$) if it holds for all x except for x in a negligible set. Denoted μ -a.e. or (a.e.). For probability measures, we say "almost surely".

Properties of Integrals:

$a, b \in \mathbb{R}^+$, $f, g \in \mathcal{E}_+$ (\mathcal{E} -measurable positive functions).

1. Positivity: $\mu(f) \geq 0$. $\mu(f) = 0 \implies f = 0$ a.e.
2. Linearity: $\mu(af + bg) = a\mu(f) + b\mu(g)$.
3. Monotonicity: If $f \leq g$ a.e., then $\mu(f) \leq \mu(g)$.

Monotone Convergence Theorem: If $f_n \rightarrow f$ from below, then $\mu(f_n) \rightarrow \mu(f)$ from below.

1. Dirac measure: using the Dirac delta $\delta_{x_0}(f) = f(x_0)$.
2. $\mu = \sum_{x \in D} m(x)\delta_x$, $D \subset E$, then $\mu(f) = \sum_{x \in D} m(x)f(x)$. Note that if E is countable, then every measure is of this form ($m(x) = \mu(\{x\})$).

Note that if E is a vector space, we can think of $\mu(f)$ as the inner product $\langle \mu, f \rangle$.

5.4 Lebesgue Integration

$E \subset \mathbb{R}^d$ a Borel set; $\mathcal{E} = \mathcal{B}(E)$. $Leb_E :=$ restriction of Leb to (E, \mathcal{E}) .

$$Leb_E(f) = \int_E Leb_E(dx)f(x) = \int_E dx f(x) = \int_E f(x)dx$$

If the Reiman integral of f exists, then Leb integral of f does as well and things are equal. However, the converse is false. Notice that if $E = [0, 1]$, $f = \mathbb{K}_{\mathbb{Q}}$, then $Leb(f) = 0$.

5.5 Integration Over a Set

$A \subset E$, $A \in \mathcal{E}$, $f \in \mathcal{E}$. Then $f\mathbb{K}_A \in \mathcal{E}$ and so $\mu(f\mathbb{K}_A) = \int f\mathbb{K}_A d\mu = \int_A f d\mu$.

5.6 Monotone Convergence Theorem

Let $\{f_n\}$ be a monotone increasing sequence of measurable positive functions. Then $\mu(\lim_{n \rightarrow \infty} f_n) = \lim_{n \rightarrow \infty} \mu(f_n)$.

Pf: $f := \lim f_n$ is well defined, so $\mu(f)$ is well defined. For all n , $f_n \leq f$, so $\mu(f_n) \leq \mu(f)$ and $\lim_n \mu(f_n) \leq \mu(f)$.

For the other direction, we want to show that $\lim_{n \rightarrow \infty} \mu(f_n) \geq \mu(d_k \circ f) \forall k$.

5.7 The Insensitivity of the Integral W.R.T Negligible Sets

Lemma: If $A \in \mathcal{E}$ is negligible, then $\int_A f d\mu = 0$ for all measurable f .

If $f = g$ a.e., then $\mu(f) = \mu(g)$.

If $f \in \mathcal{E}_+$, $\mu(f) = 0$ then $f = 0$ a.e.

5.8 Fatou's Lemma

Let $(f_n)_{n \geq 1}$ be a sequence of functions in \mathcal{E}_+ , then $\mu(\liminf f_n) \leq \liminf \mu(f_n)$. This follows from MCT (HW).

5.9 Dominated Convergence Theorem

If f_n is a sequence of functions and there exists a function g such that (a) $|f_n| \leq g \forall n$, and (b) g is integrable, then $f := \lim f_n$ (if it exists) is integrable and $\mu(f) = \lim \mu(f_n)$. This follows from Fatou's (HW).

Terminology: g dominates f_n for every n .

Cor: (Bounded Convergence Theorem) Suppose that μ is a finite measure, and $|f_n| \leq c < \infty$ (c a constant), and $f := \lim f_n$ exists. Then $\mu(f) = \lim \mu(f_n)$.

5.10 A note on these theorems...

For the majority of these theorems, there exists an always everywhere version.

5.11 Characterization of the Integral

$f \mapsto \mu(f)$, maps from \mathcal{E}_+ into $\overline{\mathbb{R}}_+$.

Theorem: Let (E, \mathcal{E}) be a measurable space. $L : \mathcal{E}_+ \rightarrow \overline{\mathbb{R}}_+$. Then there exists a unique measure μ on (E, \mathcal{E}) such that $L(f) = \mu(f)$ if and only if

- $f = 0 \implies L(f) = 0$.
- $L(af + bg) = aL(f) + bL(g)$.
- If $f_n \rightarrow f$ from below, then $L(f_n) \rightarrow L(f)$ from below.

5.12 Product Spaces

- $(E, \mathcal{E}), (F, \mathcal{F}): (E \times F, \mathcal{E} \otimes \mathcal{F})$.
- $(E, \mathcal{E}, \mu), (F, \mathcal{F}, \nu): (E \times F, \mathcal{E} \otimes \mathcal{F}, \mu \times \nu)$.
- $(\mu \times \nu)(A \times B) = \mu(A) \times \nu(B)$.

Theorem: (Fubini) Suppose that $f : E \times F \rightarrow \overline{\mathbb{R}}$ such that $\int \int_{E \times F} |f| d(\mu \times \nu) < \infty$. Then $\int \int_{E \times F} f d(\mu \times \nu) = \int_F (\int_E f(x, y) \mu(dx)) \nu(dy) = \int_E (\int_F f(x, y) \nu(dy)) \mu(dx)$.

Theorem: (Tonelli) If $f \geq 0$ then the same conclusions hold.

5.13 Absolute Continuity of Measures

Def: (E, \mathcal{E}) with measures μ and ν . We say that μ is absolutely continuous with regard to ν if $\forall A \in \mathcal{E}, \nu(A) = 0 \implies \mu(A) = 0$. Denote this by $\mu \ll \nu$.

Example: If a measure on \mathbb{R} has a density (e.g., $\mu(dx) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, the standard Gaussian measure), then it is absolutely continuous with regard to the Lebasgue measure.

Example: Discrete distributions with the same support.

Theorem: Suppose that μ is σ -finite, and that $\nu \ll \mu$. Then there exists a positive \mathcal{E} -measurable function p such that $\int_E \nu(dx) f(x) = \int_E \mu(dx) p(x) f(x), \forall f \in \mathcal{E}_+$.

Moreover, p is unique up to equivalence (if this holds for p' , then $p = p'$ a.e.).

Def: This function p is called the Radon-Nikodym derivative of ν with regard to μ . We write this as $p(x) = \frac{\nu(dx)}{\mu(dx)}(x)$, or $p = \frac{d\nu}{d\mu}$.

If we care about ν , but it is difficult to use. If $\nu \ll \mu$, then we can perform calculations using the nicer μ .

Def: μ is singular with regard to μ if there exists some set $D \in \mathcal{E}$ such that $\mu(D) = 0$ and $\nu(E \setminus D) = 0$.

6 10/2/2018

6.1 Products of Measure Spaces

$$\bigotimes_{i=1}^n (E_i, \mathcal{E}_i, \mu_i)$$

Can be seen n mutually independent random variables (i.e. n coin tosses).

How do we define a countably infinite product of measure spaces $(\bigotimes_{i=1}^{\infty} (E_i, \mathcal{E}_i, \mu_i))$.

Let \mathcal{R} be the collection of all finite dimension measurable rectangles. That is, all sets of the form $\{x \mid x_1 \in B_1, \dots, x_n \in B_n, x_{n+1} \in \mathbb{R}, x_{n+2} \in \mathbb{R}, \dots\}$ where $n \in \mathbb{N}$ and $B_i \in \mathcal{B}(\mathbb{R})$. Then $\mathcal{B}_C = \sigma(\mathcal{R})$. We define the measure as

$$\mu(\{x \mid x_1 \in B_1, \dots, x_n \in B_n, \dots\}) = \mu_1(B_1)\mu_2(B_2)\dots\mu_n(B_n)$$

Theorem: (Kolmogorov's Extension Theorem)

Suppose $\{\mu_n\}_{n \geq 1}$ is a sequence of probability measures, where μ_n is a probability measure on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ that is consistent. That is

$$\mu_{n+1}(\{x_1 \in B_1, x_2 \in B_2, \dots, x_n \in B_n, x_{n+1} \in \mathbb{R}^n\}) = \mu_n(\{x_1 \in B_1, x_2 \in B_2, \dots, x_n \in B_n\})$$

For all $n \in \mathbb{N}$ and all $B_1, B_2, \dots, B_n \in \mathcal{B}(\mathbb{R})$. Then there exists a unique probability measure \mathbb{P} on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}_C)$ such that $\mathbb{P}(\{w \mid w_1 \in B_1, \dots, w_n \in B_n\}) = \mu_n(B_1 \times B_2 \times \dots \times B_n)$.

6.2 Probability

$(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. $X : \Omega \rightarrow \mathbb{R}$ a random variable. A distribution of X is the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Note that $\mu(A) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}) = \mathbb{P}(X^{-1}A)$.

Reminder: Refresh yourself on the definitions of the common probability distributions

- Binomial
- Geometric
- Poisson
- Exponential
- Gaussian

6.3 Expected Value

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \mu(dx) = \int_{\Omega} X(\omega) \mathbb{P}(d\omega)$$

6.4 Weak Law of Large Numbers

Theorem: Suppose that X_1, X_2, \dots are i.i.d. random variables, and that $\mathbb{E}[|X_n|] < \infty$. Let $m = \mathbb{E}[X_1]$. Then $\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - m\right| \geq \epsilon\right) = 0$.

6.5 Markov's Inequality

Let X be a non-negative random variable ($X : \Omega \rightarrow \mathbb{R}_+$), and let $\lambda > 0$. Then $\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}$.

Pf: $\mathbb{P}(X \geq \lambda) = \mathbb{E}[\mathbb{1}_{\{X \geq \lambda\}}] \leq \mathbb{E}[\frac{X}{\lambda}]$.

6.6 Chebyshev's Inequality

Let X be a random variable, such that $\mathbb{E}[X^2] < \infty$. Then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\text{Var}(X)}{\lambda^2}$$

where $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Pf: $\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq \lambda^2)$. Now apply Markov's Inequality.

6.7 General Markov

Let X be a random variable and $f : \mathbb{R} \rightarrow \mathbb{R}_+$ an increasing function.

Then $\mathbb{P}(X \geq \lambda) = \mathbb{P}(f(X) \geq f(\lambda)) \leq \frac{\mathbb{E}[f(X)]}{f(\lambda)}$.

6.8 Chernoff Bound

Suppose X_1, X_2, \dots, X_n independent Bernoulli random variables, with $\mathbb{E}[X_i] = p_i$. Let $S_n = \sum_{i=1}^n X_i$, and $\mu = \sum_{i=1}^n p_i$. Then

$$\mathbb{P}(S_n \geq \mu + \lambda) \leq e^{-\frac{2\lambda^2}{n}}$$

and

$$\mathbb{P}(S_n \leq \mu - \lambda) \leq e^{-\frac{2\lambda^2}{n}}$$

In general, Chernoff provides a much tighter bound than Chebyshev.

Pf: Homework

6.9 Almost Sure Convergence

Y_1, Y_2, \dots on $(\Omega, \mathcal{F}, \mathbb{P})$.

Def: $\lim_{n \rightarrow \infty} Y_n = 0$ almost everywhere $\iff \mathbb{P}(\{\omega \mid \lim_{n \rightarrow \infty} Y_n(\omega) = 0\}) = 1$.

6.10 Strong Law of Large Numbers

Under the same conditions as before, (X_1, X_2, \dots) I.I.D., $\mathbb{E}[|X_1|] < \infty$, $m = \mathbb{E}[X_1]$ we have that $\frac{S_n}{n} \rightarrow m$ almost everywhere as $n \rightarrow \infty$.

7 10/4/18

7.1 Strong Law of Large Number

Theorem: let X_1, X_2, \dots be I.I.D. random variables with $\mathbb{E}[X_1] < \infty$ and $m = \mathbb{E}[X_1]$. Then $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow m$ almost everywhere.

In other words

$$\mathbb{P}(\{\omega \mid \lim_{n \rightarrow \infty} \sum_{i=1}^n X_i(\omega) = m\}) = 1$$

Equivalently $\frac{1}{n} \sum_{i=1}^n (X_i - m) \rightarrow 0$ almost everywhere. Without loss of generality, assume $m = 0$.

7.2 Borel-Cantelli Lemmas (Sufficient conditions for Almost Everywhere)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_i \in \mathcal{F}$ for $i \geq 1$. Define the following event

$$\limsup A_n = \bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} A_i = \{\omega \in \Omega \mid \omega \in A_n \text{ for infinitely many } n\}$$

$$\liminf A_n = \bigcup_{j=1}^{\infty} \bigcap_{i=j}^{\infty} A_i = \{\omega \in \Omega \mid \omega \in A_n \text{ for all but finitely many } n\}$$

We can think of unions and intersections of sets representing event as follows:

$$\bigcap \iff \forall \mid \bigcup \iff \exists$$

This terminology comes from $\limsup_{n \rightarrow \infty} \mathbb{1}_{A_n}(\omega) = \mathbb{1}_{\limsup A_n}(\omega)$.

Fattou: $\mathbb{P}(\liminf A_n) \leq \liminf \mathbb{P}(A_n) \leq \limsup \mathbb{P}(A_n) \leq \mathbb{P}(\limsup A_n)$.

Lemma (Borel-Cantelli I): If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup A_n) = 0$. That is, almost surely only finitely many of the events happen.

Pf: Let $\epsilon > 0$ be arbitrary. Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, there exists some N_ϵ such that $\sum_{n=N_\epsilon}^{\infty} \mathbb{P}(A_n) \leq \epsilon$. Suppose $\sum_{n=N}^{\infty} \mathbb{P}(A_n) \leq \epsilon$. Then we have

$$0 \leq \mathbb{P}(\limsup A_n) = \mathbb{P}\left(\bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} A_i\right) \leq \mathbb{P}\left(\bigcup_{i=N}^{\infty} A_i\right) \leq \sum_{i=N}^{\infty} \mathbb{P}(A_i) \leq \epsilon$$

Lemma (Borel-Cantelli II): If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = +\infty$ and A_n are mutually independent then $\mathbb{P}(\limsup A_n) = 1$.

That is, almost surely infinitely many of A_n will occur.

Pf: Want to show $\mathbb{P}(\limsup A_n) = 1$. That is equivalent to saying that $\mathbb{P}((\limsup A_n)^C) = 0$. Now,

$$(\limsup A_n)^C = \left(\bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} A_i\right)^C = \bigcup_{j=1}^{\infty} \bigcap_{i=j}^{\infty} A_i^C = \liminf A_n^C$$

Need to show that $\mathbb{P}(\bigcap_{i=j}^{\infty} A_i^C = 0)$. Fix a large number M , then $\mathbb{P}(\bigcap_{i=j}^{\infty} A_i^C) \leq \mathbb{P}(\bigcap_{i=j}^M A_i^C) = \prod_{i=j}^M \mathbb{P}(A_i^C) = \prod_{i=j}^M (1 - \mathbb{P}(A_i)) \leq \prod_{i=j}^M e^{-\mathbb{P}(A_i)} = e^{-\sum_{i=j}^M \mathbb{P}(A_i)}$. Now let $M \rightarrow \infty$ and it goes to 0.

7.3 Proof of SLLN

Theorem: Let X_1, X_2, \dots be I.I.D. random variables with $\mathcal{E}[|X_1|] < \infty$ and $\mathcal{E}[X_1] = 0$. Then $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow 0$ almost everywhere.

Step 1: Kolmogorov's inequality

Step 2: Apply inequality to show that the summable variances imply a.s. convergence.

Step 3: Kronecker's lemma and SLLN with summable variances condition

Step 4: Full SLLN proof using truncation argument

Theorem (Kolmogorov's Inequality): Let X_1, X_2, \dots, X_n be mutually independent. Assume $\mathbb{E}[X_i] = 0$ and $\sigma_i^2 = \mathbb{E}[X_i^2] < \infty$. Then for any $\lambda > 0$ we have $\mathbb{P}(\max_{1 \leq i \leq n} |X_1 + X_2 + \dots + X_n| \geq \lambda) \leq \frac{\sum_{i=1}^n \sigma_i^2}{\lambda^2}$. Note: this is a strengthening of Chebyshev.

Pf: Let $S_k = X_1 + X_2 + \dots + X_k$. Define $A = \{\omega \mid \max_{1 \leq i \leq n} |S_i| \geq \lambda\}$. Define $A_k = \{\omega \mid \max_{1 \leq i \leq k-1} |S_i| < \lambda, |S_k| \geq \lambda\}$.

Notice that $A = \bigcup_{k=1}^n A_k$, $\mathbb{1}_A = \sum_{k=1}^n \mathbb{1}_{A_k}$, $A_k \cap A_l = \emptyset$ for $k \neq l$.

$$\mathbb{P}(A) = \mathbb{E} \mathbb{1}_A = \sum_{k=1}^n \mathbb{E} \mathbb{1}_{A_k} \leq \sum_{k=1}^n \mathbb{E} \left[\frac{S_k^2}{\lambda^2} \mathbb{1}_{A_k} \right] \leq \frac{1}{\lambda^2} \sum_{k=1}^n \mathbb{E} [S_k^2 \mathbb{1}_{A_k}] + \mathbb{E} [(S_n - S_k)^2 \mathbb{1}_{A_k}]$$

8 10/16/2018

8.1 Last Time

Last time we proved the central limit theorem.

8.2 Characteristic Functions

$$\Phi_X(t) := \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]$$

Is the characteristic function of X , with X a random variable.

Properties:

1. $\Phi_X(0) = 1$
2. $\Phi_{-X}(t) = \Phi_X(-t) = \overline{\Phi_X(t)}$
3. $|\Phi_X(t)| \leq 1$
4. $t \mapsto \Phi_X(t)$ is uniformly continuous on \mathbb{R} .
5. Φ_X is positive definite. That is $\forall n$ and $\forall t_1, t_2, \dots, t_n \in \mathbb{R}$, the matrix $\{M_{i,j} = \Phi_X(t_i - t_j)\}_{i,j=1}^n$ is positive definite. That is, $\forall z_1, z_2, \dots, z_n \in \mathbb{C}$, $\sum_{i,j=1}^n z_i \Phi_X(t_i - t_j) \overline{z_j} \geq 0$.

Note: The distribution of X is symmetric around 0 if and only if Φ_X is real.

Theorem (Bochner's Theorem): Let $\Phi : \mathbb{R} \rightarrow \mathbb{C}$. Suppose that

- $\Phi(0) = 1$.
- $t \mapsto \Phi(t)$ is continuous at $t = 0$.
- Φ is positive definite.

Then Φ is the characteristic function of some random variable X . In other words, there exists a CDF F such that $\Phi(t) = \int_{-\infty}^{\infty} e^{itx} dF(x)$.

Further Properties:

- $\Phi_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = e^{ibt}\Phi_X(at)$.
- If X_1 and X_2 are independent, then $\Phi_{X_1+X_2}(t) = \Phi_{X_1}(t)\Phi_{X_2}(t)$.
- If $F' = f$ (F a CDF), then $\Phi = \hat{f}$, the Fourier transform. That is, in this case, $\Phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$.

8.3 Characteristic Functions and Moments

Theorem: Suppose that $\mathbb{E}[|X|^k] < \infty$. Then $\Phi_X \in \mathbb{C}^k$ and $\Phi_X^{(k)}(t) = \mathbb{E}[(iX)^k e^{itx}]$.

Note: \exists a random variable X such that $\mathbb{E}|X| = \infty$, but Φ_X is differentiable at $t = 0$.

Claim (Taylor Approximation Around $t = 0$): Assume that $\mathbb{E}[|X|^m] < \infty$. Then $\Phi_X(t) = \sum_{k=0}^m \mathbb{E}[X^k] \cdot \frac{(it)^k}{k!} + o(t^m)$.

Analyticity: Suppose that $\mathbb{E}[|X|^k] < \infty \forall k$ and $R^{-1} := \limsup_{m \rightarrow \infty} \left(\frac{\mathbb{E}[|X|^m]}{m!} \right)^{\frac{1}{m}} < \infty$, then Φ extends analytically to the strip $\{t + is \mid |s| < R\}$.

Example: $\mathcal{N}(0, 1)$, $R = \infty$.

8.4 Examples of Characteristic Functions

- $X \sim \text{Bernoulli}(p) : \Phi_X(t) = pe^{it} + (1-p)$.
- $X \sim \text{Bin}(n, p) : \Phi_X(t) = (pe^{it} + (1-p))^n$.
- $X \sim \text{Radenmacher} : \Phi_X(t) = \cos(t)$.
- $X \sim \text{Uni}(a, b) : \Phi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$.
- $X \sim \mathcal{N}(m, \sigma^2) : \Phi_X(t) = e^{itm - \frac{\sigma^2 t^2}{2}}$.
- $X \sim \text{Exp}(\lambda) : \Phi_X(t) = \frac{\lambda}{\lambda - it}$.
- $X \sim \text{Cauchy} : \Phi_X(t) = e^{-|t|}$.

Want to compute $\Phi_X(t)$ with $X \sim \mathcal{N}(0, 1)$. Let f_X be the density function of the standard normal distribution X . We know $\Phi_X'(t) = \mathbb{E}[iX e^{itX}] = \int_{\mathbb{R}} -x \sin(tx) f_X(x) dx = \int_{\mathbb{R}} \sin(tx) f_X'(x) dx = -t \int_{\mathbb{R}} \cos(tx) f_X(x) dx = -t \Phi_X(t)$. So,

$$\Phi_X'(t) = -t \Phi_X(t); \Phi_X(0) = 1$$

This ODE problem has a unique solution, $\Phi_X(t) = e^{-\frac{t^2}{2}}$.

8.5 Inversions

Theorem (Levy's Inversion Theorem): Suppose that X has CDF F_X and characteristic function Φ_X . For every real numbers $a < b$ and t . Let

$$\Psi_{a,b}(t) := \frac{1}{2\pi} \int_a^b e^{-itu} du = \frac{e^{-tb} - e^{-ita}}{-i2\pi t}$$

Then $\lim_{T \rightarrow \infty} \int_{-T}^T \Psi_{a,b}(t) \Phi_X(t) dt = \frac{1}{2} [F_X(b) + F_X(b-)] - \frac{1}{2} [F_X(a) + F_X(a-)]$.

In particular, if a and b are continuity points of F_X , then the limit is $F_X(b) - F_X(a)$.

Furthermore, if $\int_{\mathbb{R}} |\Phi_X(t)| dt < \infty$ then X has the following bounded and continuous probability density function:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \Phi_X(t) dt$$

This is a special case of the standard Fourier inversion. The formula here is an integrated version which holds also in the absence of a density.

Corollary: If $\Phi_X(t) = \Phi_Y(t) \forall t \in \mathbb{R}$, then X and Y have the same distribution.

Proof left as an exercise.

9 10/18/2018

9.1 Last Time

The "Law" of X , written $\mathcal{L}(X)$, is in one to one correspondence with Φ_X .

9.2 Levy's Continuity Theorem

Let $\{F_n\}_{n \geq 1}$ be a sequence of CDFs on \mathbb{R} and let $\{\Phi_n\}_{n \geq 1}$ be the corresponding characteristic functions.

- If $F_n \rightarrow F$ then $\Phi_n(t) \rightarrow \Phi(t)$ for all $t \in \mathbb{R}$ (pointwise).
- Suppose that for every $t \in \mathbb{R}$, the limit $\lim_{n \rightarrow \infty} \Phi_n(t)$ exists and denote it by $\Phi(t)$. Suppose that Φ is continuous at $t = 0$. Then \exists CDF F such that $\Phi(t) = \int e^{itx} dF(x)$ and $F_n \rightarrow F$.

9.3 Central Limit Theorem Proof

Theorem: Suppose X_1, X_2, \dots are I.I.D. such that $\mathbb{E}[X_1^2] < \infty$. Assume $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$. Let $S_n = X_1 + \dots + X_n$. Then $\frac{S_n}{\sqrt{n}} \rightarrow \mathcal{N}(0, 1)$.

Proof: $\Phi_{X_1}(t) = \mathbb{E}[e^{itX_1}]$. As $t \rightarrow 0$,

$$\Phi_{X_1}(t) = 1 + \mathbb{E}[X_1] \frac{it}{1!} + \mathbb{E}[X_1^2] \frac{(it)^2}{2!} + o(t^2) = 1 - \frac{t^2}{2 + o(t^2)}$$

where $f(t) = o(t^2)$ if $\lim_{t \rightarrow 0} \frac{f(t)}{t^2} = 0$. Now,

$$\begin{aligned}
& \Phi_{\frac{S_n}{\sqrt{n}}} \\
&= \mathbb{E}[e^{it(\frac{S_n}{\sqrt{n}})}] \\
&= \mathbb{E}[e^{(\frac{it}{\sqrt{n}})(X_1 + \dots + X_n)}] \\
&= \mathbb{E}[e^{(\frac{it}{\sqrt{n}})X_1}] \dots \mathbb{E}[e^{(\frac{it}{\sqrt{n}})X_n}] \\
&= (\Phi_{X_1}(\frac{t}{\sqrt{n}}))^n \\
&= (1 - \frac{(\frac{t}{\sqrt{n}})^2}{2} + o((\frac{t}{\sqrt{n}})^2))^n \\
&= (1 - \frac{t^2}{2} + o(\frac{t^2}{2}))^n \\
&\rightarrow e^{-\frac{t^2}{2}}
\end{aligned}$$

9.4 Some Comments on Weak Convergence

Def: A sequence of probability measures $\{\mu_n\}_{n \geq 1}$ on \mathbb{R} is tight if $\forall \epsilon > 0$, $\exists K < \infty$ such that $\forall n \geq 1$, $\mu_n([-K, K]) > 1 - \epsilon$. Intuitively, mass doesn't "disappear to infinity" if $\mu_n = \mu$.

Counterexamples:

- $\mu_n = \delta_n$
- $\mu_n : \mathcal{N}(0, \sigma_n^2 = n)$
- $Uni(-n, n)$.

Def: Let $\{\mu_n\}_{n \geq 1}$ be a probability measure on a complete metric space S . This sequence is tight if $\forall \epsilon > 0$, $\exists K \subset S$ such that $\forall n \geq 1$, $\mu_n(K) > 1 - \epsilon$.

Claim: If $F_n \rightarrow F$, then $\{F_n\}_{n \geq 1}$ is tight. Proof left as an exercise.

Theorem (Helly): If $\{F_n\}_{n \geq 1}$ are tight, then $\exists \{n_k\}_{k \geq 1}$ (subsequence) and a CDF F such that $F_{n_k} \rightarrow F$.

Theorem (Prohorov): Let S be a complete separable metric space. If $\{\mu_n\}_{n \geq 1}$ are tight, then \exists a weakly convergent subsequence $\{\mu_{n_k}\}_{k \geq 1}$.

9.5 Generaly Strategy for Proving Weak Convergence (Prohorov)

1. Show that $\{\mu_n\}_{n \geq 1}$ is tight.
2. Identify the limit μ .
3. Show uniqueness of the limit.

9.6 Stochastic Processes

- Index set \mathbb{T} representing time (EX: $\mathbb{R}, \mathbb{R}_+, [a, b], \mathbb{Z}, \mathbb{Z}_+, \dots$)
- State space S (locally compact complete metric space)

$$t \mapsto X_t; X : \mathbb{T} \times \Omega \rightarrow S$$

Where $\forall t \in \mathbb{T}; X(t, \cdot) : \Omega \rightarrow S$ is measurable.

1. $t \in \mathbb{T}, X_t : \Omega \rightarrow S$ (marginal)
2. $\omega \in \Omega$ fixed, $X(\cdot, \omega) : \mathbb{T} \rightarrow S$ a (random) function.
3. $X : \Omega \rightarrow S^{\mathbb{T}} = \{\text{functions } \mathbb{T} \rightarrow S\}$. A random variable taking values in function space.

Examples:

- $\{X_n\}_{n \geq 1}$ I.I.D.
- $S_n = \sum_{i=1}^n Y_i$ where $\{Y_i\}_{n \geq 1}$ are I.I.D.
- Random walks on a graph.

9.7 Markov Processes and Markov Chains

Markov Process: A stochastic process where the distribution of the future given the past and present only depends on the present.

Discrete Time and Discrete Space: In this case, Markov Processes are called Markov Chains.

Def: X_0, X_1, \dots is a Markov Chain on a discrete state space S if $\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \forall n \geq 1, \forall x_0, x_1, \dots, x_n \in S$. That is, the distribution of X_n depends only on the result of X_{n-1} .

Notation: $X_0^n = (X_0, X_1, \dots, X_n)$.

Notice that, in general,

$$\mathbb{P}(X_0^n = x_0^n) = \mathbb{P}(X_n = x_n \mid x_0^{n-1} = x_0^{n-1}) \cdot \mathbb{P}(X_0^{n-1} = x_0^{n-1}) = \dots = \prod_{i=1}^n \mathbb{P}(X_i = x_i \mid X_0^{i-1} = x_0^{i-1})$$

So for Markov chains,

$$\mathbb{P}(X_0^n = x_0^n) = \mathbb{P}(X_0 = x_0) \cdot \prod_{i=1}^n \mathbb{P}(X_i = x_i \mid X_{i-1} = x_{i-1})$$

Notation:

$$\mathbb{P}(X_n = y \mid X_{n-1} = x) = P_{x,y}(n)$$

This is called the transition probabilities index. That is, the probability of going from x to y at time n . This gives the transition probability matrix at time n :

$$P(n) = \{P_{x,y}(b)\}_{x,y \in S}$$

Often the transition probabilities are not a function of n . In that case, we say that the Markov chain is time-homogeneous, and write

$$P = \{P_{x,y}\}_{x,y \in S}$$

9.8 Properties of Transition Probabilities Matrix

Notice that for a Markov chain, $\mathbb{P}(X_0^n = x_0^n) = \mathbb{P}(X_0 = x_0) \cdot \prod_{i=1}^n P_{x_{i-1}, x_i} P(i)$. So, two things determine a Markov chain

- The initial distribution X_0 .
- Transition probabilities.

So

- $0 \leq P_{x,y} \leq 1$.
- $\sum_{y \in S} P_{x,y} = 1 \quad \forall x \in S$.

10 10/25/2018

10.1 Markov Chains

For the time being we will consider finite state spaces.

Say we want to understand the probability that we end up at a particular state at timestep n . We can do so by taking powers of the transition probability matrix. That is, $\mathbb{P}(X_n = y \mid X_0 = x) = (P^n)_{x,y}$.

Pf:

This is true for $n = 1$ by definition. For $n > 1$, we can use conditioning

$$\begin{aligned} \mathbb{P}(X_n = y \mid X_0 = x) &= \sum_{z \in S} \mathbb{P}(X_n = y, X_{n-1} = z \mid X_0 = x) \\ &= \sum_{z \in S} \mathbb{P}(X_n = y \mid X_{n-1} = z, X_0 = x) \\ &= \sum_{z \in S} \mathbb{P}(X_n = y \mid X_{n-1} = z) \mathbb{P}(X_{n-1} = z \mid X_0 = x) \\ &= \sum_{z \in S} P_{zy} (P^{n-1})_{xz} = (P^n)_{xy} \end{aligned}$$

Now, think of what P defines. P is a matrix,

- It acts on column vectors to its right. Think of these column vectors as functions.
- It acts on row vectors to the left. Think of these as probability measures.

Suppose $f : S \rightarrow \mathbb{R}$. Then $(P^n f)(x) = \mathbb{E}[f(X_n) \mid X_0 = x]$.

Pf: $(Pf)(x) = \sum_{y \in S} P_{xy} f(y) = \sum_{y \in S} \mathbb{P}(X_1 = y \mid X_0 = x) f(y) = \mathbb{E}[f(X_1) \mid X_0 = x]$.

Now suppose $\mu : S \rightarrow \mathbb{R}$ is a measure. Then $(\mu P^n)(x) = \mathbb{P}_\mu(X_n = x)$ here \mathbb{P}_μ indicates that the initial distribution is μ .

$$(\mu P^n)(x) = \sum_{y \in S} \mu(y) (P^n)_{yx} = \sum_{y \in S} \mu(y) \mathbb{P}(X_n = x \mid X_0 = y)$$

Note: P acts to the right on functions.

$$l^\infty(S) = \{f : S \rightarrow \mathbb{R} \mid \|f\|_\infty < \infty\}$$

Properties:

- Keeps positivity: If $f \geq 0$, then $Pf \geq 0$.
- P is a contraction: $\|P\|_{\infty, \infty} \leq 1$.
- $\|Pf\|_\infty = \max_{x \in S} |(Pf)(x)| = \max_{x \in S} |\sum_y P_{xy} f(y)| \leq \max_{x \in S} \sum_y P_{xy} |f(y)| \leq \max_{x \in S} \sum_y P_{xy} \|f\|_\infty = \|f\|_\infty$.
- $P1 = 1$ (That is, vector of all 1s).

Note: P acts to the left on measures.

$$l^1(S) = \{\mu : S \rightarrow \mathbb{R} \mid \|\mu\|_1 < \infty\} = \sum_{x \in S} |\mu(x)|.$$

1. Keeps positivity.
2. Is a contraction. $\|\mu P\|_1 \leq \|\mu\|_1$.
3. $\exists \mu$ such that $\mu P = \mu$.

Def: If $\mu P = \mu$, and μ is a probability measure, then μ is a stationary distribution.

Claim: If $\mu P = \mu$, then also $|\mu|P = |\mu|$ where $|\mu|(x) = |\mu(x)|$.

Corollary: If $\mu P = \mu$, then $\mu^+ P = \mu^+$, $\mu^- P = \mu^-$.

Corollary: If S is finite, then a stationary probability measure with respect to P always exists.

Pf:

$$(|\mu|P)(x) = \sum_{y \in S} |\mu(y)| P_{yx} \geq \left| \sum_{y \in S} \mu(y) P_{yx} \right| = |\mu(x)|$$

Suppose that $\exists x \in S$ such that $(|\mu|P)(x) > |\mu(x)|$. Then $\| |\mu|P \|_1 = \sum_{y \in S} (|\mu|P)(y) > \sum_{y \in S} |\mu(y)| = \| |\mu| \|_1$. This contradicts the contraction property.

10.2 Graphs

Let $G = (V, E)$ be a directed graph with $(i, j) \in E$ (indicating a change of state from i to j) if and only if $P_{ij} > 0$.

Def: $A \subset S$ is closed if $\mathbb{P}(A \rightarrow A^C) = 0$. That is, once the state is in A , it cannot enter a state outside of A .

Note: \emptyset, S are closed. If A, B are closed, then $A \cap B$ and $A \cup B$ are also closed.

Def: $A \subset S$, $\bar{A} = \bigcap_{B \supset A; B \text{ closed}} B$. This is the closure of A .

Def: Suppose that $A = \bar{A}$ and $A \neq \emptyset$. Then A is said to be irreducible if $\forall B \subset A, B \neq \emptyset$ we have $\bar{B} = A$. Otherwise, A is called reducible.

Def: $x \in S$ is an absorbing state if $\{x\}$ is irreducible.

Def: $x \in S$ is an inessential state if it is not part of any irreducible component. Otherwise, the state is essential.

Often we will focus on irreducible Markov chains, because over a long time scale, only the irreducible components of any Markov chain are relevant.

Periodicity:

Def: The period of $x \in S$ is the greatest common divisor of all walks that start in x and come back to x .

Claim: Suppose $x, y \in S$ such that there exists some path $x \rightarrow y$ and a path $y \rightarrow x$. Then $\text{per}(x) = \text{per}(y)$. Proof left as an exercise.

Corollary: The period of every state in an irreducible component is the same.

Def: The period of an irreducible component is equal to the period of its states.

Def: An irreducible Markov chain (a chain composed of a single irreducible component) has a period equal to the period of its states.

Fact: The number of stationary measures of a Markov chain is equal to the number of irreducible components of that chain.

Def: P is ergodic if it is irreducible and is aperiodic.

Def: If a state, component, or chain has a period of 1, we call it aperiodic.

Theorem: Suppose that X_0, X_1, \dots is an ergodic Markov chain on a finite state space S with transition probability matrix P . For all $x \in S$, if $X_0 = x$, $X_n \rightarrow \Pi$ as $n \rightarrow \infty$ where Π is the unique stationary distribution.