Measure Theoretic Probability

Ethan Berger Anderes

October 31, 2016

Abstract

The majority of these notes are heavily inspired by, and closely follow, a set of class notes developed by Michael Wichura in the Statistics Department at the University of Chicago. These notes had a profound impact on me and to generations of PhD students. All the credit for the organization, exposition and clarity must goes to him. The mistakes, wherever they may be, are entirely due to myself.

Note: These notes are currently work in progress. Many of the theorems are stated without proof (but will be proved in class). I plan to sequentially add proofs and examples as time permits. There are some shaded regions in the notes, which indicated the material is still under construction. Moreover, there are some sections which have no content. These will hopefully be added as time goes on.

Contents

Part I
Measure

1 Borel's normal number theorem

Definition 1 (**Borel field).** *Let $\mathcal{B}_0((0,1])$ denote the class of finite (possibly empty) disjoint unions of intervals of the form $(a,b] \subset (0,1]$.*

Definition 2. *Let P be a probability assignment on $\mathcal{B}_0((0,1])$ such that $P[(a,b]] = b-a$ for all $0 \le a \le b \le 1$ and extended to all of $\mathcal{B}_0((0,1])$ using the identity $P[A \cup B] = P[A] + P[B]$ whenever A, B are disjoint sets in $\mathcal{B}_0((0,1])$.*

Theorem 1. *P is well defined.*

Definition 3. *For each $\omega \in (0,1]$, let $d_k(\omega)$ denote the k^{th} nonterminating binary digit of ω. Let $z_k(\omega) := 2d_k(\omega) - 1$ and $s_n(\omega) := \sum_{k=1}^n z_k(\omega) \equiv$ excess of heads in n tosses.*

Theorem 2 (**WLLN).** *For all $\epsilon > 0$,*

$$\lim_{n\to\infty} P\Big[\big\{\omega \in (0,1] : |s_n(\omega)/n| \ge \epsilon\big\}\Big] = 0. \tag{1}$$

Proof. Notice first that any event or bet based on the values of $z_1(\omega), \dots, z_n(\omega)$ must be a disjoint union of dyadic intervals of the form $(\frac{k-1}{2^n}, \frac{k}{2^n}]$. Therefore $\big\{\omega \in (0,1] : |s_n(\omega)/n| \ge \epsilon\big\} \in \mathcal{B}_0((0,1])$ and the left hand side of (1) is well defined. Also notice

$$\int_0^1 z_k(\omega) z_j(\omega)\, d\omega = \begin{cases} 1 & \text{when } k = j \\ 0 & \text{when } k \ne j. \end{cases}$$

This implies $\int_0^1 s_n^2(\omega)\, d\omega = \int_0^1 \sum_{k,j=1}^n z_k(\omega) z_j(\omega)\, d\omega = n$ which gives

$$\begin{aligned} n = \int_0^1 s_n^2(\omega)\, d\omega &\ge \int_{|s_n/n| \ge \epsilon} s_n^2(\omega)\, d\omega \\ &\ge \int_{|s_n/n|\ge\epsilon} n^2\epsilon^2\, d\omega \ge n^2\epsilon^2 P\big[|s_n/n| \ge \epsilon\big] \end{aligned}$$

Therefore $P\big[|s_n/n| \ge \epsilon\big] \le 1/(n\epsilon^2) \to 0$ as $n \to \infty$. $\square$

Definition 4. *The set of* **normal numbers** *in $(0,1]$ is defined as*

$$\begin{aligned} N &:= \{\omega \in (0,1] : \lim_{n\to\infty} s_n(\omega)/n = 0\} \\ &= \{\omega \in (0,1] : \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{k=1}^n d_k(\omega) = \tfrac{1}{2}\}. \end{aligned}$$

The set of abnormal numbers is defined as $A := (0,1] - N$.

Definition 5 (**Negligible set).** *A subset $B \subset (0,1]$ is said to be* **negligible** *if for all $\epsilon > 0$, there exists $\mathcal{B}_0((0,1])$-sets $B_1, B_2, \dots$ such that*

$$B \subset \bigcup_{k=1}^\infty B_k \quad \text{and} \quad \sum_{k=1}^\infty P[B_k] \le \epsilon.$$

Theorem 3 (**Borel's normal number theorem, i.e. the SLLN for coin flips).** *The set of abnormal numbers, A, is negligible.*

Proof. Let $\epsilon_k \downarrow 0$ as $k \to \infty$. Then

$$\begin{aligned} \{\omega \colon |\tfrac{s_{k^2}(\omega)}{k^2}| < \epsilon_k \text{ for all large } k\} & \\ \subset \{\omega \colon \lim_k \tfrac{s_{k^2}(\omega)}{k^2} = 0\} & \\ \subset \underbrace{\{\omega \colon \lim_n \tfrac{s_n(\omega)}{n} = 0\}}_{=N} & \end{aligned} \tag{2}$$

To see why (2) holds assume $\lim_k s_{k^2}(\omega)/k^2 = 0$ for ω and notice that

$$\begin{aligned} \Big|\frac{s_n}{n}\Big| = \frac{|s_n|}{(\sqrt{n})^2} \le \frac{|s_n|}{\lfloor\sqrt{n}\rfloor^2} &\le \frac{s_{\lfloor\sqrt{n}\rfloor^2}}{\lfloor\sqrt{n}\rfloor^2} + \frac{|s_n - s_{\lfloor\sqrt{n}\rfloor^2}|}{\lfloor\sqrt{n}\rfloor^2} \\ &\le \frac{s_{\lfloor\sqrt{n}\rfloor^2}}{\lfloor\sqrt{n}\rfloor^2} + \sum_{k=\lfloor\sqrt{n}\rfloor^2+1}^n \frac{|z_k|}{\lfloor\sqrt{n}\rfloor^2} \\ &= \frac{s_{\lfloor\sqrt{n}\rfloor^2}}{\lfloor\sqrt{n}\rfloor^2} + \frac{(n - \lfloor\sqrt{n}\rfloor^2)}{\lfloor\sqrt{n}\rfloor^2} \longrightarrow 0 \end{aligned}$$

since $n - \lfloor\sqrt{n}\rfloor^2 \le (\lfloor\sqrt{n}\rfloor + 1)^2 - \lfloor\sqrt{n}\rfloor^2 = 1 + 2\lfloor\sqrt{n}\rfloor$. Now by (2) we have that

$$\begin{aligned} A = N^c \subset \{\omega \colon |\tfrac{s_{k^2}(\omega)}{k^2}| \ge \epsilon_k \text{ for infinitely many } k\} & \\ \subset \bigcup_{k=j}^\infty \underbrace{\{\omega \colon |\tfrac{s_{k^2}(\omega)}{k^2}| \ge \epsilon_k\}}_{=:B_k}, \quad \text{for any } j & \end{aligned}$$

where $B_k \in B_0^{(0,1])}$. By the proof of the WLLN we have

$$P[B_k] \le \frac{1}{k^2\epsilon_k^2} = \frac{1}{k^{3/2}}$$

when $\epsilon_k := k^{-1/4}$. Therefore $\sum_{k=1}^n P[B_k] < \infty$ and hence $\sum_{k=j}^n P[B_k] \to 0$ as $j \to \infty$. Hence A is negligible. $\square$

Exercise 1. *Using just calculus and ideas from this section, show that*

$$M(t) := \int_0^1 e^{ts_n(\omega)}\, d\omega = \Big(\frac{e^t + e^{-t}}{2}\Big)^n \tag{3}$$

for each $t \in \mathbb{R}$. By differentiating with respect to t, show that $\int_0^1 s_n(\omega)\, d\omega = M'(0) = 0$ and $\int_0^1 s_n^2(\omega)\, d\omega = M''(0) = n$.

Exercise 2. *Show that*

$$P\big[|s_n/n| \ge \epsilon\big] \le 2e^{-n\epsilon^2/2}$$

for each $\epsilon > 0$.

Hint: Use (3) in conjunction with the inequality $(e^x + e^{-x})/2 \le \exp(x^2/2)$ which holds (why?) for all $x \in \mathbb{R}$.

2 Classes of sets

Definition 6. *Ω denotes the* **sample space**. *Subsets of Ω are called* **events** *and 2^Ω denotes the power set of Ω (i.e. the class of all subsets of Ω).*

Definition 7 (field). *A collection of events $\mathcal{F} \subset 2^\Omega$ is a* **field** *if*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$
3. $A, B \in \mathcal{F} \Longrightarrow A \cup B \in \mathcal{F}$.

Definition 8 (σ-field). *A collection of events $\mathcal{F} \subset 2^\Omega$ is a* **σ-field** *if*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots \in \mathcal{F} \Longrightarrow \bigcup_{k=1}^\infty A_k \in \mathcal{F}$.

Definition 9 (λ-system). *A collection of events $\mathcal{F} \subset 2^\Omega$ is called a* **λ-system** *if*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$
3. $\underbrace{A_1, A_2, \ldots}_{\text{all disjoint}} \in \mathcal{F} \Longrightarrow \bigcup_{k=1}^\infty A_k \in \mathcal{F}$.

Notice that the only reason we require $\Omega \in \mathcal{F}$ in the definitions above is to force the class $\mathcal{F}$ be non-empty. We could just as well have changed the requirement $\Omega \in \mathcal{F}$ to the statment that there exists some $A \in \mathcal{F}$. One of the reasons it is traditional to put the assumption $\Omega \in \mathcal{F}$ is that the definition of a probability measure will require $P(\Omega) = 1$. Therefore, it makes things more clear if we explicitly claim that $\Omega \in \mathcal{F}$, but otherwise its superfluous.

Definition 10 (π-system). *A collection of events $\mathcal{P} \subset 2^\Omega$ is called a* **π-system** *if*

1. $A, B \in \mathcal{P} \Longrightarrow A \cap B \in \mathcal{P}$.

Definition 11 ($A_n \uparrow A$). *Let $A_1, A_2, \ldots$ and A be events of Ω. Then we write $\lim_n\uparrow A_n = A$ (or $A_n \uparrow A$) if*

1. $A_1 \subset A_2 \subset \cdots$
2. $A = \bigcup_{k=1}^\infty A_k$.

Definition 12 ($A_n \downarrow A$). *Let $A_1, A_2, \ldots$ and A be events of Ω. Then we write $\lim_n\downarrow A_n = A$ (or $A_n \downarrow A$) if*

1. $A_1 \supset A_2 \supset \cdots$
2. $A = \bigcap_{k=1}^\infty A_k$.

Definition 13 (monotone class). *A collection of events $\mathcal{M} \subset 2^\Omega$ is a* **monotone class** *if*

1. $A_1, A_2, \ldots \in \mathcal{M}$ *and* $A_n \uparrow A \Longrightarrow A \in \mathcal{M}$
2. $A_1, A_2, \ldots \in \mathcal{M}$ *and* $A_n \downarrow A \Longrightarrow A \in \mathcal{M}$.

Theorem 4 ($\sigma = \lambda + \pi = \mathscr{M} + f$).

$$\mathcal{F} \text{ is a } \sigma\text{-field} \Longleftrightarrow \mathcal{F} \text{ is a field and a monotone class} \tag{4}$$
$$\Longleftrightarrow \mathcal{F} \text{ is a } \lambda\text{-system and a } \pi\text{-system.} \tag{5}$$

Proof. (*show (5)*) Notice the direction ($\Longrightarrow$) is trivial. To show the other direction suppose $\mathcal{F}$ is a λ-system and a π-system. We need to show $\mathcal{F}$ is a σ-field. Notice $\Omega \in \mathcal{F}$ is trivial by λ-system properties. Also $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ is trivial by λ-system properties. To show $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_{k=1}^\infty \in \mathcal{F}$ one uses a common trick for turning a non-disjoint union into a disjoint union.

$$\begin{aligned}\bigcup_{k=1}^\infty A_k &= \bigcup_{k=1}^\infty \underbrace{A_k - (A_1 \cup \cdots \cup A_{k-1})}_{disjoint} \\ &= \bigcup_{k=1}^\infty A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c\end{aligned}$$

Now $A_k^c \in \mathcal{F}$ by λ-system properties and hence $A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c \in \mathcal{F}$ by π-system properties. Therefore $\cup_{k=1}^\infty A_k$ can be written as a disjoint union of events from $\mathcal{F}$. Therefore $\cup_{k=1}^\infty A_k \in \mathcal{F}$ by λ-system properties as was to be shown.

(*show (4)*) Just as in the proof of (5) the only non-trivial thing to show is that when $\mathcal{F}$ is a field and a monotone class this implies that $\mathcal{F}$ is closed under countable union. Indeed if $A_1, A_2, \ldots \in \mathcal{F}$ then

$$\bigcup_{k=1}^\infty A_k = \lim_n\uparrow \bigcup_{k=1}^n A_k$$

where $\bigcup_{k=1}^n A_k \in \mathcal{F}$ by the field properties and therefore $\bigcup_{k=1}^\infty A_k \in \mathcal{F}$ by the monotone class properties. $\square$

2.1 Generators

Theorem 5 (field generated by $\mathcal{C}$). *Let $\mathcal{C} \subset 2^\Omega$. Then*

$$f\langle\mathcal{C}\rangle := \bigcap_{\substack{\mathcal{F} \text{ is a field} \\ \mathcal{C} \subset \mathcal{F}}} \mathcal{F}$$

is a field (which contains $\mathcal{C}$).

Theorem 6 (σ-field generated by $\mathcal{C}$). *Let $\mathcal{C} \subset 2^\Omega$. Then*

$$\sigma\langle\mathcal{C}\rangle := \bigcap_{\substack{\mathcal{F} \text{ is a } \sigma\text{-field} \\ \mathcal{C} \subset \mathcal{F}}} \mathcal{F}$$

is a σ-field (which contains $\mathcal{C}$).

Theorem 7 (**monotone class generated by $\mathcal{C}$**)**.** *Let $\mathcal{C} \subset 2^\Omega$. Then*

$$\mathscr{M}\langle\mathcal{C}\rangle := \bigcap_{\substack{\mathcal{M} \text{ is a monotone class} \\ \mathcal{C} \subset \mathcal{M}}} \mathcal{M}$$

is a monotone class (which contains $\mathcal{C}$).

Theorem 8 (**λ-system generated by $\mathcal{C}$**)**.** *Let $\mathcal{C} \subset 2^\Omega$. Then*

$$\lambda\langle\mathcal{C}\rangle := \bigcap_{\substack{\mathcal{L} \text{ is a } \lambda\text{-system} \\ \mathcal{C} \subset \mathcal{L}}} \mathcal{L}$$

is a λ-system (which contains $\mathcal{C}$).

Theorem 9 (**Good sets**)**.** *Let $\mathcal{C}$ and $\mathcal{G}$ be two collections of subsets of Ω. If*

- *$\mathcal{C} \subset \mathcal{G}$;*
- *$\mathcal{G}$ is a σ-field*

Then $\sigma\langle\mathcal{C}\rangle \subset \mathcal{G}$.

Theorem 10 (**Restricted generators**)**.** *Let Ω be a sample space and $\mathcal{C}$ be a class of subsets of Ω. If $\Omega_0 \subset \Omega$ then*

$$\underbrace{\sigma\langle\mathcal{C} \cap \Omega_0\rangle}_{\substack{\sigma\text{-field} \\ \text{on } \Omega_0}} = \underbrace{\sigma\langle\mathcal{C}\rangle}_{\substack{\sigma\text{-field} \\ \text{on } \Omega}} \cap \Omega_0.$$

Proof. (*Show $\sigma\langle\mathcal{C} \cap \Omega_0\rangle \subset \sigma\langle\mathcal{C}\rangle \cap \Omega_0$*) This easily follows by *good sets* since clearly $\mathcal{C} \cap \Omega_0 \subset \sigma\langle\mathcal{C}\rangle \cap \Omega_0$ and Exercise 3 shows that $\sigma\langle\mathcal{C}\rangle \cap \Omega_0$ is a σ-field.

(*Show $\sigma\langle\mathcal{C}\rangle \cap \Omega_0 \subset \sigma\langle\mathcal{C} \cap \Omega_0\rangle$*) Notice that this inclusion is equivalent to the statement that for every $A \in \sigma\langle\mathcal{C}\rangle$, $A \cap \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle$. To show this let

$$\mathcal{G} := \{A \subset \Omega : A \cap \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle\}.$$

It will then be sufficient to show the following four bullets and then use good sets to conclude $\sigma\langle\mathcal{C}\rangle \subset \mathcal{G}$.
•($\mathcal{C} \subset \mathcal{G}$) $A \in \mathcal{C} \Longrightarrow A \cap \Omega_0 \in \mathcal{C} \cap \Omega_0 \subset \sigma\langle\mathcal{C} \cap \Omega_0\rangle$.
•($\Omega \in \mathcal{G}$)

$$\begin{aligned}\Omega_0 \subset \Omega &\Longrightarrow \Omega \cap \Omega_0 = \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\qquad\qquad \text{since a } \sigma\text{-field on } \Omega_0 \text{ must contain } \Omega_0 \\ &\Longrightarrow \Omega \in \mathcal{G}.\end{aligned}$$

•($A \in \mathcal{G} \Longrightarrow A^c \in \mathcal{G}$) Notice that A^c denotes complementation within Ω. Now

$$\begin{aligned}A \in \mathcal{G} &\Longrightarrow A \cap \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\Longrightarrow \underbrace{\Omega_0 - A \cap \Omega_0}_{\text{complement in } \Omega_0} \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\Longrightarrow \underbrace{\Omega_0 \cap (A^c \cup \Omega_0^c)}_{=A^c \cap \Omega_0} \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\Longrightarrow A^c \in \mathcal{G}.\end{aligned}$$

•($A_1, A_2, \ldots \in \mathcal{G} \Longrightarrow \bigcup_k A_k \in \mathcal{G}$)

$$\begin{aligned}A_1, A_2, \ldots \in \mathcal{G} &\Longrightarrow A_k \cap \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle, \forall k \\ &\Longrightarrow \bigcup_k (A_k \cap \Omega_0) \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\Longrightarrow \Big(\bigcup_k A_k\Big) \cap \Omega_0 \in \sigma\langle\mathcal{C} \cap \Omega_0\rangle \\ &\Longrightarrow \bigcup_k A_k \in \mathcal{G}.\end{aligned}$$

□

Theorem 11 (**Halmos's monotone class theorem**)**.** *If $\mathcal{F}_0$ is a field then $\mathscr{M}\langle\mathcal{F}_0\rangle = \sigma\langle\mathcal{F}_0\rangle$.*

Theorem 12 (**Dynkin's $\pi - \lambda$ theorem**)**.** *If $\mathcal{P}$ is a π-system then $\lambda\langle\mathcal{P}\rangle = \sigma\langle\mathcal{P}\rangle$.*

Proof. This proof uses *good sets* all over the place. First notice that $\lambda\langle\mathcal{P}\rangle \subset \sigma\langle\mathcal{P}\rangle$ follows directly from *good sets* since $\mathcal{P} \subset \sigma\langle\mathcal{P}\rangle$ and clearly $\sigma\langle\mathcal{P}\rangle$ is also a λ-system. Therefore we only need to show $\sigma\langle\mathcal{P}\rangle \subset \lambda\langle\mathcal{P}\rangle$.

Each statement below gives a sufficient condition to establish that $\sigma\langle\mathcal{P}\rangle \subset \lambda\langle\mathcal{P}\rangle$. They are given in reverse dependency order to make it easier to follow the train of reasoning.

$$\begin{gathered}\sigma\langle\mathcal{P}\rangle \subset \lambda\langle\mathcal{P}\rangle \\ \Uparrow \\ \lambda\langle\mathcal{P}\rangle \text{ is a } \sigma\text{-field} \\ \Uparrow \\ \lambda\langle\mathcal{P}\rangle \text{ is a } \pi\text{-system} \\ \Uparrow \\ \forall A, B \in \lambda\langle\mathcal{P}\rangle \text{ one has } A \cap B \in \lambda\langle\mathcal{P}\rangle \\ \Uparrow \\ \forall A \in \lambda\langle\mathcal{P}\rangle \text{ one has } \lambda\langle\mathcal{P}\rangle \subset \mathcal{G}_A \text{ where} \\ \mathcal{G}_A := \big\{B \subset \Omega \colon A \cap B \in \lambda\langle\mathcal{P}\rangle\big\} \\ \Uparrow \\ \forall A \in \lambda\langle\mathcal{P}\rangle, \mathcal{P} \subset \mathcal{G}_A \text{ and } \mathcal{G}_A \text{ is a } \lambda\text{-system.}\end{gathered}$$

The last statement above is what we show. Notice, first, that

$$A \in \mathcal{G}_B \Longleftrightarrow A \cap B \in \lambda\langle\mathcal{P}\rangle \Longleftrightarrow B \in \mathcal{G}_A. \tag{6}$$

In particular if $A \cap B \in \lambda\langle\mathcal{P}\rangle$ then one has that both $A \in \mathcal{G}_B$ and $B \in \mathcal{G}_A$.

(*Case 1: show $\mathcal{P} \subset \mathcal{G}_A$ and $\mathcal{G}_A$ is a λ-system when $A \in \mathcal{P}$*)
•($\mathcal{P} \subset \mathcal{G}_A$) If $B \in \mathcal{P}$ then $A \cap B \in \mathcal{P}$ by the π-system properties of $\mathcal{P}$. Therefore $B \in \mathcal{G}_A$.
•($\Omega \in \mathcal{G}_A$) This follows since $A \cap \Omega = A \in \mathcal{P}$.
•($B \in \mathcal{G}_A \Longrightarrow B^c \in \mathcal{G}_A$)

$$\begin{aligned}B \in \mathcal{G}_A &\Longrightarrow A \cap B \in \lambda\langle\mathcal{P}\rangle \\ &\Longrightarrow A - A \cap B \in \lambda\langle\mathcal{P}\rangle, \quad \lambda\text{-system properties} \\ &\Longrightarrow A - B \in \lambda\langle\mathcal{P}\rangle\end{aligned}$$

$$\Longrightarrow A \cap B^c \in \lambda\langle \mathcal{P} \rangle$$
$$\Longrightarrow B^c \in \mathcal{G}_A.$$

•(disjoint $B_1, B_2, \ldots \in \mathcal{G}_A \Longrightarrow \cup_k B_k \in \mathcal{G}_A$) Notice $A \cap \bigcup_{k=1}^\infty B_k = \bigcup_{k=1}^\infty (A \cap B_k)$. The $A \cap B_k$'s are disjoint if the B_k's are too. Since B_k's are in $\mathcal{G}_A$, by assumption, we must have $A \cap B_k \in \lambda\langle \mathcal{P} \rangle$. Therefore $\bigcup_{k=1}^\infty (A \cap B_k) \in \lambda\langle \mathcal{P} \rangle$ by λ-system properties. Therefore $A \cap \bigcup_{k=1}^\infty B_k \in \mathcal{G}_A$.

(*Case 2: show $\mathcal{P} \subset \mathcal{G}_A$ and $\mathcal{G}_A$ is a λ-system when $A \in \lambda\langle \mathcal{P} \rangle$*)
•($\mathcal{P} \subset \mathcal{G}_A$) The only reason we established Case 1 was to proof this part of Case 2. Indeed, Case 1 establishes that when $A \in \mathcal{P}$ we have that $\lambda\langle \mathcal{P} \rangle \subset \mathcal{G}_A$ by *good sets*. Changing names gives $\lambda\langle \mathcal{P} \rangle \subset \mathcal{G}_B$ whenever $B \in \mathcal{P}$. Now

$$\begin{aligned} B \in \mathcal{P} &\Longrightarrow \lambda\langle \mathcal{P} \rangle \subset \mathcal{G}_B, \quad \text{from Case 1} \\ &\Longrightarrow A \in \mathcal{G}_B \\ &\Longrightarrow B \in \mathcal{G}_A, \quad \text{by (6).} \end{aligned}$$

•($\Omega \in \mathcal{G}_A$) Same as in Case 1.
•($B \in \mathcal{G}_A \Longrightarrow B^c \in \mathcal{G}_A$) Same as in Case 1.
•(disjoint $B_1, B_2, \ldots \in \mathcal{G}_A \Longrightarrow \cup_k B_k \in \mathcal{G}_A$) Same proof as in Case 1.

□

Theorem 13 (Good sets, take 2). *Let $\mathcal{P}$ and $\mathcal{G}$ be two collections of subsets of Ω. If*

- *$\mathcal{P} \subset \mathcal{G}$;*
- *$\mathcal{P}$ is a π-system;*
- *$\mathcal{G}$ is a λ-system*

Then $\sigma\langle \mathcal{P} \rangle \subset \mathcal{G}$.

Exercise 3. *Suppose $\mathcal{F}$ is a σ-field on Ω and let Ω_0 be any subset of Ω (not necessarily in $\mathcal{F}$). Prove that $\mathcal{F} \cap \Omega_0 := \{F \cap \Omega_0 : F \in \mathcal{F}\}$ is a σ-field on Ω_0.*

Exercise 4. *Prove Halmos's monotone class theorem. (Hint: To show $\sigma\langle \mathcal{F}_0 \rangle \subset \mathscr{M}\langle \mathcal{F}_0 \rangle$ notice that it will be sufficient to show that $\mathscr{M}\langle \mathcal{F}_0 \rangle$ is a field (why?); then to show that $\mathscr{M}\langle \mathcal{F}_0 \rangle$ is a field start by showing it is closed under complementation, then under intersection.)*

Exercise 5. *For any non-empty class $\mathcal{A} \subset 2^\Omega$, if*

$$\begin{aligned} \mathcal{C} &:= \text{the collection of } \mathcal{A} \text{ sets and their complements} \\ \mathcal{I} &:= \text{the collection of finite intersections of } \mathcal{C} \text{ sets} \\ \mathcal{U} &:= \text{the collection of finite unions of } \mathcal{I} \text{ sets.} \end{aligned}$$

then $f\langle \mathcal{A} \rangle = \mathcal{U}$. Hint: first show $\mathcal{U}$ is closed under intersections, then complements.

Definition 14 (Semi-ring with unit). *A collection of events $\mathcal{A} \subset 2^\Omega$ is called a* **semi-ring with unit** *if*

1. $\Omega \in \mathcal{A}$
2. $A, B \in \mathcal{A} \Longrightarrow A \cap B \in \mathcal{A}$
3. *If $A \in \mathcal{A}$ then A^c is a finite disjoint union of $\mathcal{A}$-sets*

Exercise 6. *Suppose $\mathcal{A} \subset 2^\Omega$ is a semi-ring with unit. Let $\mathscr{D}$ denote the class of finite disjoint unions of $\mathcal{A}$-sets. Show $f\langle \mathcal{A} \rangle = \mathscr{D}$. Hint: first show $\mathscr{D}$ is closed intersections, then complements.*

Exercise 7. *Show that $\mathcal{B}_0((0,1])$ from Definition 1 is a field and coincides with $f\langle (a,b] : 0 \le a \le b \le 1 \rangle$.*

Exercise 8. *Let $\Omega = \mathbb{R}$. Show that $f\langle (-\infty, a] : -\infty < a < \infty \rangle$ is the the set of finite (possibly empty) disjoint unions of intervals of the form $(-\infty, b]$, (a, ∞) and $(a, b]$ for finite $a < b$. (Hint: change the generators a bit to apply exercise 6.)*

Exercise 9. *Let $\mathcal{A} \subset 2^\Omega$ be a countable collection of Ω sets. Show that $f\langle \mathcal{A} \rangle$ is a countable collection of Ω sets.*

Exercise 10. *Let $\mathcal{L}$ be a collection of subsets of Ω. Show that $\mathcal{L}$ is a λ-system if and only if $\mathcal{L}$ satisfies the following three conditions*

1. $\Omega \in \mathcal{L}$
2. *If $A - B \in \mathcal{L}$ whenever $B \subset A$ and $A, B \in \mathcal{L}$*
3. *$A_1, A_2, \ldots \in \mathcal{L}$ and $A_n \uparrow A \Longrightarrow A \in \mathcal{L}$.*

2.2 Borel σ-fields

Borel σ-fields are used throught the whole theory of measure and integration. In this section we go into detail treatment of these fields. The main story is that Borel σ-fields have many equivenlent generators. Different generators are useful for proving different things. For example the Borel σ-field on $(0,1]^d$ as generated by the field of finite disjoint unions of rectangles is useful for constructing Lebesque measure. To specidfy uniqueness of a measure on $\mathbb{R}^d$ with a particular property it is often useful to consider the Broel field on $\mathbb{R}^d$ to be $\sigma\big\langle (-\infty, c_1] \times \cdots \times (-\infty, c_d] : -\infty < c_k < \infty \big\rangle$ the generators of which form a π-system.

Definition 15 (Metric space Borel σ-field: $\mathcal{B}(\Omega)$). *Suppose Ω forms a metric space with some metric $d : \Omega \times \Omega \to [0, \infty]$. A set $A \subset \Omega$ is said to be* **open** *if for each $x \in A$, there exists an $\epsilon > 0$ such that the open ball $\{y \in \Omega : d(x,y) < \epsilon\}$ is contained in A. The* **Borel σ-field of Ω** *(with respect to metric d), denoted $\mathcal{B}(\Omega)$, is defined as the σ-field generated by the open sets.*

The above definition immediately allows us to define the Borel σ-fields $\mathcal{B}(\mathbb{R}^d)$ and $\mathcal{B}((0,1]^d)$. To define $\mathcal{B}(\bar{\mathbb{R}}^d)$, where $\bar{\mathbb{R}} := [-\infty, \infty]$ we use the metric given by $d(x,y) := |\tau(x) - \tau(y)|$ where

$$\tau(x) := \begin{cases} \frac{x}{1+|x|} & \text{when } |x| < \infty; \\ 1 & \text{when } x = \infty; \\ -1 & \text{when } x = -\infty. \end{cases} \tag{7}$$

Theorem 14 (Borel restrictions). *Let Ω be a metric space and $\Omega_o \subset \Omega$. Then the Borel σ-field $\mathcal{B}(\Omega_o)$, which is constructed using the induced metric on Ω, satisfies*

$$\mathcal{B}(\Omega_o) = \mathcal{B}(\Omega) \cap \Omega_o$$

If, in addition, $\Omega_o \in \mathcal{B}(\Omega)$ then $\mathcal{B}(\Omega_o) = \{B \in \mathcal{B}(\Omega) : B \subset \Omega_o\}$.

Proof. (*Show* $\mathcal{B}(\Omega_o) = \mathcal{B}(\Omega) \cap \Omega_o$) Let

$$\begin{aligned}\mathcal{G} &:= \text{open subsets of } \Omega \\ \mathcal{G}_o &:= \text{open subsets of } \Omega_o\end{aligned}$$

Notice that Theorem 2.30 in Rudin (Principles in Mathematical Analysis) shows that

$$\mathcal{G}_o = \mathcal{G} \cap \Omega_o.$$

This implies

$$\begin{aligned}\mathcal{B}(\Omega_o) = \sigma\langle \mathcal{G}_o \rangle &= \sigma\langle \mathcal{G} \cap \Omega_o \rangle \\ &= \sigma\langle \mathcal{G} \rangle \cap \Omega_o, \text{ by Theorem 10} \\ &= \mathcal{B}(\Omega) \cap \Omega_o.\end{aligned}$$

(*Show* $\mathcal{B}(\Omega) \cap \Omega_o = \{B \in \mathcal{B}(\Omega) : B \subset \Omega_o\}$ *whenever* $\Omega_o \in \mathcal{B}(\Omega)$). To see '$\supset$' suppose $B \subset \Omega_o$ and $B \in \mathcal{B}(\Omega)$. Then $B = B \cap \Omega_o \in \mathcal{B}(\Omega) \cap \Omega_o$. To see '$\subset$' let $B \in \mathcal{B}(\Omega) \cap \Omega_o$ so that $B = \tilde{B} \cap \Omega_o$ where $\tilde{B} \in \mathcal{B}(\Omega)$. Since $\Omega_o \subset \Omega$ we have $B \in \mathcal{B}(\Omega)$ and $B \subset \Omega_o$. □

Theorem 15 (Non-exhaustive list of useful Borel generators).

$$\begin{aligned}\mathcal{B}(\mathbb{R}^d) &= \sigma\big\langle (-\infty, c_1] \times \cdots \times (-\infty, c_d] : -\infty < c_k < \infty \big\rangle \\ &= \sigma\big\langle \textit{open balls of } \mathbb{R}^d \big\rangle \\ &= \sigma\big\langle \textit{open subsets of } \mathbb{R}^d \big\rangle \\ &= \sigma\big\langle \textit{closed subsets of } \mathbb{R}^d \big\rangle \\ &= \sigma\big\langle \textit{compact subsets of } \mathbb{R}^d \big\rangle \\ &= \sigma\big\langle \textit{rectangles in } \mathbb{R}^d \big\rangle \\ &= \sigma\big\langle \textit{cylinders } \mathbb{R}^d \big\rangle\end{aligned}$$

$$\begin{aligned}\mathcal{B}((0,1]) &= \sigma\langle \mathcal{B}_0((0,1]) \rangle \\ &= \sigma\big\langle (a,b] : 0 \le a \le b \le 1 \big\rangle \\ &= \sigma\big\langle (a,b) : 0 < a < b < 1 \big\rangle \\ &= \sigma\big\langle [a,b] : 0 < a < b < 1 \big\rangle \\ &= \sigma\big\langle (0,a] : 0 < a < 1 \big\rangle \\ &= \sigma\big\langle \textit{open subsets of } (0,1] \big\rangle \\ &= \sigma\big\langle \textit{closed subsets of } (0,1] \big\rangle\end{aligned}$$

$$\begin{aligned}\mathcal{B}((0,1]^d) &= \mathcal{B}(\mathbb{R}^d) \cap (0,1]^d \\ &= \{B \in \mathcal{B}(\mathbb{R}^d) : B \subset (0,1]^d\} \\ &= \sigma\big\langle (a_1,b_1] \times \cdots \times (a_d,b_d] : 0 \le a_k < b_k \le 1 \big\rangle \\ &= \sigma\big\langle \mathcal{B}_0((0,1]^d) \big\rangle.\end{aligned}$$

where $\mathcal{B}_0((0,1]^d) := f\big\langle (a_1,b_1] \times \cdots \times (a_d,b_d] : 0 \le a_k < b_k \le 1 \big\rangle$ is the Borel field on $(0,1]^d$ which equals the finite (possibly empty) disjoint union of rectangles from $\{(a_1,b_1] \times \cdots \times (a_d,b_d] : 0 \le a_k < b_k \le 1\}$

I would venture to say that one of the most important results above is that $\mathcal{B}((0,1]^d) = \sigma\big\langle \mathcal{B}_0((0,1]^d) \big\rangle$ where $\mathcal{B}_0((0,1]^d)$ is the field of finite (possibly empty) disjoint union of rectangles. This characterization allows one to construct probabilities on $\mathcal{B}_0((0,1]^d)$, then use the Carathéodory Extension Theorem to extend this to a full probability model on $\mathcal{B}((0,1]^d)$.

Notice that most of the equalities in Theorem 15 are shown using the good sets principle. In particular, to show that $\sigma\langle \mathcal{A}_1 \rangle = \sigma\langle \mathcal{A}_2 \rangle$ one simply needs to establish that $\mathcal{A}_1 \subset \sigma\langle \mathcal{A}_2 \rangle$ (which implies that $\sigma\langle \mathcal{A}_1 \rangle \subset \sigma\langle \mathcal{A}_2 \rangle$ by "good sets") and $\mathcal{A}_2 \subset \sigma\langle \mathcal{A}_1 \rangle$ (which implies that $\sigma\langle \mathcal{A}_2 \rangle \subset \sigma\langle \mathcal{A}_1 \rangle$ by "good sets").

Proof. I will only show one of these equalities. The rest follow by similar arguments. To show

$$\sigma\big\langle (a,b] : 0 < a < b < 1 \big\rangle = \sigma\big\langle (a,b) : 0 < a < b < 1 \big\rangle$$

it will be sufficient to show the following two statements for any arbitrary $0 < a_0 < b_0 < 1$.

•(*Show* $(a_0,b_0] \in \sigma\langle (a,b) : 0 < a < b < 1 \rangle$) This is follows from the identity

$$(a_0, b_0] = \bigcap_{n=1}^{\infty} (a_0, b_0 + n^{-1}).$$

•(*Show* $(a_0,b_0) \in \sigma\langle (a,b] : 0 < a < b < 1 \rangle$) This is follows from the identity

$$(a_0, b_0) = \bigcup_{n=1}^{\infty} (a_0, b_0 - n^{-1}].$$

□

The sets in the Borel σ-field are extremely rich. In fact, it is hard to show that there are sets which are not in $\mathcal{B}(\mathbb{R})$. The easiest way to find such a set is to use properties of Lebseque measure which we will construct later in the notes. Therefore, we postpone a discussion of such sets until we have Lebseque measure at our disposal. For the reminder of this section we give some examples of sets which *are* in the Borel σ-fields on Euclidean space.

Example 1. *The set of normal and abnormal numbers are in $\mathcal{B}((0,1])$.*

Example 2. *All countable, co-countable (i.e. complements of countable sets), and perfect subsets of $(0,1]$ are in $\mathcal{B}((0,1])$. In particular, the collection of irrational numbers in $(0,1]$ is a Borel set.*

Exercise 11. *Show the Cantor set is an uncountable set in* $\mathcal{B}((0,1])$ *(a nice way to see that it is uncountable is to work with a base-3 digit characterization of the Cantor set).*

Exercise 12. *Let* Ω *be a metric space with distance function* d*.* Ω *is said to be* **separable** *if there exists a countable* $\Omega_0 \subset \Omega$ *which is dense in* Ω *(i.e., every point of* Ω *is a limit of some sequence of points of* Ω_0*).*

1. *Show that* $\sigma\langle \textit{open balls in } \Omega\rangle \subset \mathcal{B}(\Omega)$*.*

2. *Show that* $\sigma\langle \textit{open balls in } \Omega\rangle = \mathcal{B}(\Omega)$ *if* Ω *is separable.*

3. *Show that* $\Omega = \bar{\mathbb{R}}$ *is separable with the metric defined with (7) and conclude that* $\sigma\langle \textit{open balls in } \bar{\mathbb{R}}\rangle = \mathcal{B}(\bar{\mathbb{R}})$*.*

Exercise 13.

1. *Show that* $\mathcal{B}(\bar{\mathbb{R}})$ *is generated by the sets of the form* $[-\infty, a]$ *for* $-\infty < a < \infty$*, and also by sets of the form* $[-\infty, a)$ *for* $-\infty < a < \infty$*.*

2. *Show that* $\mathcal{B}(\bar{\mathbb{R}})$ *is not generated by sets of the form* $(-\infty, a)$ *for* $-\infty < a < \infty$*. (Hint: find a* σ*-field which contains the intervals* $(-\infty, a)$ *but which is strictly smaller than* $\mathcal{B}(\bar{\mathbb{R}})$ *).*

3 Measures

Definition 16 (**finitely additive probability**). *If $\mathcal{F}_0$ is a field on Ω, then $P : \mathcal{F}_0 \to [0,1]$ is said to be a* **finitely additive probability on** $\mathcal{F}_0$ *if*

1. $P[\Omega] = 1$
2. $P[A \cup B] = P[A] + P[B]$
   for all disjoint $A, B \in \mathcal{F}_0$.

Definition 17 (**probability measure**). *If $\mathcal{F}_0$ is a field on Ω, then $P : \mathcal{F}_0 \to [0,1]$ is said to be a* **probability measure on** $\mathcal{F}_0$ *if*

1. $P(\Omega) = 1$
2. $P\big(\bigcup_{k=1}^\infty A_k\big) = \sum_{k=1}^\infty P(A_k)$
   for all disjoint $A_1, A_2, \ldots \in \mathcal{F}_0$ such that $\bigcup_{k=1}^\infty A_k \in \mathcal{F}_0$.

Definition 18 (**Measure**). *If $\mathcal{F}_0$ is a field of Ω-sets, then $\mu : \mathcal{F}_0 \to [0,\infty]$ is a measure if*

1. $\mu(\varnothing) = 0$
2. $\mu\big(\bigcup_{k=1}^\infty A_k\big) = \sum_{k=1}^\infty \mu(A_k)$
   for all disjoint $A_1, A_2, \ldots \in \mathcal{F}_0$ such that $\bigcup_{k=1}^\infty A_k \in \mathcal{F}_0$.

Definition 19 (**Measurable space**). *If $\mathcal{F}$ is a σ-field on Ω then the pair $(\Omega, \mathcal{F})$ is called a measurable space.*

Definition 20 (**Measure/Probability space**). *If μ is a measure on the measurable space $(\Omega, \mathcal{F})$ then the triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space. If, in addition, μ is a probability measure then the triple $(\Omega, \mathcal{F}, \mu)$ is called a probability space.*

Definition 21 (**Finite and σ-finite**). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

- *If $\mu(\Omega) < \infty$ then μ is said to be a finite measure;*
- *If $\mu(\Omega) = \infty$ then μ is said to be an infinite measure;*
- *If there exists $\mathcal{F}$-sets $A_1, A_2, \ldots$ such that $\Omega = \cup_{k=1}^\infty A_k$ and $\mu(A_k) < \infty$ then μ is said to be a σ-finite measure;*
- *If $\mathscr{A} \subset \mathcal{F}$ such that there exists $\mathscr{A}$-sets $A_1, A_2, \ldots$ such that $\Omega = \cup_{k=1}^\infty A_k$ and $\mu(A_k) < \infty$ then μ is said to be σ-finite on $\mathscr{A}$.*

Note that every probability measure is σ-finite. Most of the basic rules of probability follow from finitely additive properties.

Theorem 16 (**Basic properties of probability measures**). *Suppose P is a probability measure on field $\mathcal{F}_0$ of Ω subsets. Then each of the following statements hold for all $\mathcal{F}_0$-sets $A, A_1, \ldots, B, B_1, \ldots$*

1. $P[A^c] = 1 - P[A]$*;*
2. $P[\varnothing] = 0$*;*
3. *If $A \subset B$ then $P[B - A] = P[B] - P[A]$;*
4. **(Increasing)** *If $A \subset B$ then $P[A] \le P[B]$;*
5. **(Inclusion-exclusion)** $P[A \cup B] = P[A] + P[B] - P[A \cap B]$*;*
6. **(Finite additivity)** *If A_k's are disjoint then $P\big[\bigcup_{k=1}^n A_k\big] = \sum_{k=1}^n P[A_k]$;*
7. **(Finite sub-additivity)** $P\big[\bigcup_{k=1}^n A_k\big] \le \sum_{k=1}^n P[A_k]$*;*
8. **(Approximation)** *If $A_k \subset B_k$ then $P\big[\bigcup_{k=1}^n B_k\big] - P\big[\bigcup_{k=1}^n A_k\big] \le \sum_{k=1}^n P[B_k - A_k]$ and $P\big[\bigcap_{k=1}^n B_k\big] - P\big[\bigcap_{k=1}^n A_k\big] \le \sum_{k=1}^n P[B_k - A_k]$;*
9. *(***Continuous from below***)*
   If $A_n \uparrow A \in \mathcal{F}_0$ then $P[A_n] \uparrow P[A]$;
10. *(***Continuous from above***)*
    If $A_n \downarrow A \in \mathcal{F}_0$ then $P[A_n] \downarrow P[A]$;
11. *(***Countable sub-additivity***)*
    If $\cup_{k=1}^\infty A_k \in \mathcal{F}_0$ then $P\big[\sum_{k=1}^\infty A_k\big] \le \sum_{k=1}^\infty P[A_k]$;

Proof. •(*Show item 2*) If $A \subset B$ then $B = A \cup (B - A)$ is a disjoint union. Therefore $P[B] = P[A] + P[B - A]$ which implies $P[B - A] = P[B] - P[A]$.
•(*Show item 3*) Use $0 \le P[B - A] = P[B] - P[A]$.
•(*Show item 4*) Use the fact that $A \cup B$ can be written as a disjoint union $A \cup (B - A \cap B)$ to get that

$$\begin{aligned} P[A \cup B] &= P[A] + P[B - A \cap B] \\ &= P[A] + P[B] - P[A \cap B]. \end{aligned}$$

•(*Show item 5*) Use induction.
•(*Show item 6*) Use induction and inclusion exclusion.
•(*Show item 8*) For the first equation use the fact that $\cup_k B_k - \cup_k A_k$ and $\cap_k B_k - \cap_k A_k$ are covered by $\cup_k (B_k - A_k)$ and then apply sub-additivity.
•(*Show item 9*) Assume $A_1, A_2, \ldots \in \mathcal{F}_0$ and $A_n \uparrow A \in \mathcal{F}_0$. Then

$$\begin{aligned} P[A_n] &= P\big[\textstyle\bigcup_{k=1}^n A_k\big], \qquad \text{since } A_1 \subset A_2 \subset \cdots \\ &= P\big[\textstyle\bigcup_{k=1}^n \underbrace{A_k - A_{k-1}}_{\text{disjoint}}\big] \\ &= \textstyle\sum_{k=1}^n P\big[A_k - A_{k-1}\big] \\ &\uparrow \textstyle\sum_{k=1}^\infty P\big[A_k - A_{k-1}\big] \\ &= P\big[\textstyle\bigcup_{k=1}^\infty A_k - A_{k-1}\big], \text{ by item 1.} \\ &= P[A] \end{aligned}$$

•(*Show item 10*) Use the fact that $A_n \downarrow A \iff A_n^c \uparrow A^c$.
•(*Show item 11*)

$$\begin{aligned} P\Big[\bigcup_{k=1}^\infty A_k\Big] &= P\Big[\lim_n\uparrow \bigcup_{k=1}^n A_k\Big] \\ &= \lim_n\uparrow P\Big[\bigcup_{k=1}^n A_k\Big], \text{ continuity from below} \end{aligned}$$

$\leq \lim_n\uparrow \sum_{k=1}^n P\big[A_k\big]$, finite sub-additivity
$= \sum_{k=1}^\infty P\big[A_k\big]$.

□

Theorem 17 (Basic properties of measures). *Suppose μ is a probability measure on field $\mathcal{F}_0$ of Ω subsets. Then each of the following statements hold for all $\mathcal{F}_0$-sets $A, A_1, \ldots, B, B_1, \ldots$*

1. $\mu[A^c] = \mu[\Omega] - \mu[A]$ *if* $\mu[\Omega] < \infty$;
2. *If* $A \subset B$ *and* $\mu[A] < \infty$ *then* $\mu[B - A] = \mu[B] - \mu[A]$;
3. **(Increasing)** *If* $A \subset B$ *then* $\mu[A] \leq \mu[B]$;
4. **(Inclusion-exclusion)** $\mu[A \cup B] = \mu[A] + \mu[B - A \cap B]$ *and* $\mu[A \cup B] = \mu[A] + \mu[B] - \mu[A \cap B]$ *if* $\mu[A \cap B] < \infty$;
5. **(Finite additivity)** *If A_k's are disjoint then* $\mu\big[\bigcup_{k=1}^n A_k\big] = \sum_{k=1}^n \mu[A_k]$;
6. **(Finite sub-additivity)** $\mu\big[\bigcup_{k=1}^n A_k\big] \leq \sum_{k=1}^n \mu[A_k]$.
7. *(***Continuous from below***)* *If* $A_n \uparrow A \in \mathcal{F}_0$ *then* $\mu[A_n] \uparrow \mu[A]$;
8. *(***Continuous from above***)* *If* $A_n \downarrow A \in \mathcal{F}_0$ *and there exists an n such that* $\mu[A_n] < \infty$ *then* $\mu[A_n] \downarrow \mu[A]$.
9. *(***Countable sub-additivity***)* *If* $\cup_{k=1}^\infty A_k \in \mathcal{F}_0$ *then* $\mu\big[\sum_{k=1}^\infty A_k\big] \leq \sum_{k=1}^\infty \mu[A_k]$;

The following theorem tells us that a consequences of the countable additivity which do not follow from finite additivity.

Theorem 18 (Countable additivity equivalence). *Let $P : \mathcal{F}_0 \to [0,1]$ be a finitely additive probability on a field $\mathcal{F}_0$. Then the following statements are equivalent:*

1. *P is a probability measure;*
2. *P is continuous from below;*
3. *P is continuous from above;*
4. *(***Continuous from above at $\varnothing$***) If whenever $A_1, A_2, \ldots \in \mathcal{F}_0$ and $A_n \downarrow \varnothing$ then $P(A_n) \downarrow 0$.*

Proof. By a previous theorem already have *1.* $\Longrightarrow$ *2.* $\Longrightarrow$ *3.* $\Longrightarrow$ *4.*

(*4.* $\Longrightarrow$ *3.*) Suppose P is continuous from above at $\varnothing$. Now suppose $A_1, A_2, \ldots \in \mathcal{F}_0$ and $A_n \downarrow A \in \mathcal{F}_0$. Now to show item 3 notice

$$\begin{aligned} A_n \downarrow A &\Longrightarrow A_n - A \downarrow \varnothing \\ &\Longrightarrow P[A_n] - P[A] = P[A_n - A] \downarrow 0, \text{ by item 4.} \\ &\Longrightarrow P[A_n] \downarrow P[A] \end{aligned}$$

(*2.* $\Longleftrightarrow$ *3.*) Use the fact that $A_n \uparrow A \Longleftrightarrow A_n^c \downarrow A^c$ along with the identity $P[A] = 1 - P[A^c]$.

(*2.* $\Longrightarrow$ *1.*) The only thing to show is countable additivity over disjoint sets which stay in the field. In paritcular suppose $A_1, A_2, \ldots$ are disjoint $\mathcal{F}_0$-sets such that $\bigcup_{k=1}^\infty A_k \in \mathcal{F}_0$. Then

$$P\Big[\bigcup_{k=1}^\infty A_k\Big] = P\Big[\lim_n\uparrow \bigcup_{k=1}^n A_k\Big] = \lim_n\uparrow P\Big[\bigcup_{k=1}^n A_k\Big]$$

where the last equality follows by assuming item 2.

□

Theorem 19 (Uniqueness for probability measures). *Let $\mathcal{P}$ be a collection of subsets of Ω. If P and Q are two probability measures on $(\Omega, \sigma\langle\mathcal{P}\rangle)$ such that*

1. *P and Q agree on $\mathcal{P}$;*
2. *$\mathcal{P}$ is a π-system,*

then P and Q agree on all of $\sigma\langle\mathcal{P}\rangle$.

Proof. This is our first use of Dynkin's $\pi - \lambda$ theorem which allows us to extend the good sets principle. In particular, define the good sets as follows:

$$\mathcal{G} := \{A \subset \Omega \colon Q[A] = P[A]\}. \tag{8}$$

Dynkin's $\pi - \lambda$ theorem says that $\sigma\langle\mathcal{P}\rangle = \lambda\langle\mathcal{P}\rangle$ since $\mathcal{P}$ is a π-system. Therefore to show $\sigma\langle\mathcal{P}\rangle = \lambda\langle\mathcal{P}\rangle \subset \mathcal{G}$ we just show that $\mathcal{G}$ is a λ-system and invoke *good sets*.

•($\Omega \in \mathcal{G}$) This is trivial since $Q[\Omega] = 1$ and $P[\Omega] = 1$ by properties probability measures.
•($A \in \mathcal{G} \Longrightarrow A^c \in \mathcal{G}$)

$$\begin{aligned} A \in \mathcal{G} &\Longrightarrow Q[A] = P[A] \\ &\Longrightarrow 1 - Q[A^c] = 1 - P[A^c] \\ &\Longrightarrow Q[A^c] = P[A^c] \\ &\Longrightarrow A^c \in \mathcal{G}. \end{aligned}$$

•(*disjoint* $A_1, A_2 \in \mathcal{G} \Longrightarrow \bigcup_{k=1}^\infty A_k \in \mathcal{G}$)

$$\begin{aligned} \underbrace{A_1, A_2, \ldots}_{\text{disjoint}} \in \mathcal{G} &\Longrightarrow Q[A_k] = P[A_k],\ \forall k \\ &\Longrightarrow \underbrace{\sum_{k=1}^\infty Q[A_k]}_{=Q[\cup_{k=1}^\infty A_k]} = \underbrace{\sum_{k=1}^\infty P[A_k]}_{=P[\cup_{k=1}^\infty A_k]} \\ &\Longrightarrow \bigcup_{k=1}^\infty \in \mathcal{G}. \end{aligned}$$

Notice that this last statement could not be proved if the A_k's were not disjoint. This illustrates the necessity of Dynkin's $\pi - \lambda$ theorem.

□

Theorem 20 (**Uniqueness for measures**). *If μ_1 and μ_2 are measures on $(\Omega, \sigma\langle\mathcal{P}\rangle)$ such that*

1. *μ_1 and μ_2 agree on $\mathcal{P}$;*
2. *$\mathcal{P}$ is a π-system;*
3. *μ_1 and μ_2 are σ-finite on $\mathcal{P}$,*

then μ_1 and μ_2 agree on all of $\sigma\langle\mathcal{P}\rangle$.

Definition 22 (**μ-null and μ-neg**). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

- *A set $A \in \mathcal{F}$ is said to be μ-null if $\mu(A) = 0$.*
- *A set $A \in 2^\Omega$ is said to be μ-negligible if there exists a μ-null set $B \in \mathcal{F}$ such that $A \subset B$.*

Definition 23 (**Complete**). *A measure space $(\Omega, \mathcal{F}, \mu)$ is said to be complete if all the μ-negligible sets belong to $\mathcal{F}$.*

Definition 24 (**Complete**). *A probability space $(\Omega, \mathcal{F}, P)$ is said to be* **complete** *if all the P-negligible sets belong to $\mathcal{F}$.*

Theorem 21 (**The completion $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$**). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\mathcal{N}_\mu$ be the collection of μ-negligible sets. Then*

- *$\bar{\mathcal{F}} := \sigma\langle\mathcal{F}, \mathcal{N}_\mu\rangle = \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}_\mu\}$;*
- *The set function $\bar{\mu}$ on $\bar{\mathcal{F}}$ defined by $\bar{\mu}(F \cup N) = \mu(F)$ for $F \in \mathcal{F}$ and $N \in \mathcal{N}_\mu$ is the unique extension of μ to a measure on $(\Omega, \bar{\mathcal{F}})$;*
- *The measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is complete.*

The triple $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is called the completion of $(\Omega, \mathcal{F}, \mu)$.

Exercise 14. *Suppose that μ_1 and μ_2 are measures on $\sigma\langle\mathcal{F}_0\rangle$ generated by a class $\mathcal{F}_0$. Suppose also that the inequality*

$$\mu_1(A) \leq \mu_2(A) \tag{9}$$

holds for all A in $\mathcal{F}_0$. (a) Show that if $\mathcal{F}_0$ is a field and μ_1 and μ_2 are σ-finite on $\mathcal{F}_0$, then (9) holds for all $A \in \sigma\langle\mathcal{F}_0\rangle$. (b) Show by examples that (9) can fail for some $A \in \sigma\langle\mathcal{F}_0\rangle$ if: $\mathcal{F}_0$ is a field but μ_1 and μ_2 are only σ-finite overall, not σ-finite on $\mathcal{F}_0$; or if μ_1 and μ_2 are σ-finite on $\mathcal{F}_0$, but $\mathcal{F}_0$ is only a π-system. Hint: for (a) first treat the case where μ_2 is finite.

The following exercises shows that it doesn't matter what order we sum an infinite number of positive terms. This is useful for rigorously showing the measure axioms for counting measure on any Ω. Notice that if there are negative terms, order does matter.

Exercise 15. *Let I be an infinite set and let f be a function from I to $[0, \infty]$. The sum over f over I is defined as*

$$\sum_{i\in I} f(i) := \sup\Big\{\sum_{i\in H} f(i) : \ H \subset I, \ H \text{ is finite}\Big\}.$$

Show that: (a) for any partition $I = \bigcup_{k\in K} I_k$ of I into nonempty disjoint subsets I_k,

$$\sum_{i\in I} f(i) = \sum_{k\in K}\Big(\sum_{i\in I_k} f(i)\Big)$$

and (b) if I is countable, then

$$\sum_{i\in I} f(i) = \lim_{n\to\infty}\sum_{m=1}^{n} f(i_m)$$

for any enumeration $i_1, i_2, \ldots$ of the points in I.

Exercise 16. *Let $\Omega = \mathbb{R}$ and $\mathcal{B}_0(\mathbb{R}) := f\langle(-\infty, a] : -\infty < a < \infty\rangle$ be the Borel field of $\mathbb{R}$. Let P be a finitely additive probability on $\mathcal{B}_0(\mathbb{R})$. Show that P is a probability measure on $\mathcal{B}_0(\mathbb{R})$ if and only if the function defined by $F(x) := P((-\infty, x])$ is non-decreasing, right-continuous and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.*

3.1 Carathéodory Extension Theorem

Section Assumption. *For the remainder of this section P_0 denotes a probability measure on $\mathcal{F}_0$, where $\mathcal{F}_0$ is a field on Ω. Also let $\mathcal{F}^\uparrow, \mathcal{F}^\downarrow, \bar{\mathcal{F}}, P^\uparrow, P^\downarrow, P^*, P_*, \bar{P}$ be defined as follows*

- *$\mathcal{F}^\uparrow := \{\lim_k\!\uparrow A_k : A_k \in \mathcal{F}_0\}$*
- *$\mathcal{F}^\downarrow := \{\lim_k\!\downarrow A_k : A_k \in \mathcal{F}_0\}$*
- *$P^\uparrow(\lim_k\!\uparrow A_k) := \lim_k P_0(A_k)$ when $\lim_k\!\uparrow A_k \in \mathcal{F}^\uparrow$*
- *$P^\downarrow(\lim_k\!\downarrow A_k) := \lim_k P_0(A_k)$ when $\lim_k\!\downarrow A_k \in \mathcal{F}^\downarrow$*
- *$P^*(A) := \inf\{P^\uparrow(A^\uparrow) : A \subset A^\uparrow \in \mathcal{F}^\uparrow\}$ when $A \in 2^\Omega$*
- *$P_*(A) := \sup\{P^\downarrow(A^\downarrow) : A \supset A^\downarrow \in \mathcal{F}^\downarrow\}$ when $A \in 2^\Omega$*
- *$\bar{\mathcal{F}} := \{A \in 2^\Omega : P^*(A) = P_*(A)\}$*
- *$\bar{P}(A) := P^*(A) = P_*(A)$ when $A \in \bar{\mathcal{F}}$*

Theorem 22. *$P^\uparrow$, $P^\downarrow$, P^* and P_* are all well defined. Moreover, $(2^\Omega, P^*)$ is an extension of $(\mathcal{F}^\uparrow, P^\uparrow)$ which is an extension of $(\mathcal{F}_0, P_0)$ (and similarly for $(2^\Omega, P_*)$ and $(\mathcal{F}^\downarrow, P^\downarrow)$).*

Proof. (*Show $P^\uparrow$ is well defined*) Notice that if $A_n \uparrow A$ then $\lim_n P_0[A_n]$ exists by monotonicity and boundedness of $P_0[A_n]$. Therefore we just need to show that $\lim_n P_0[A_n] = \lim_n P_0[B_n]$ whenever $\lim_n\!\uparrow A_n = \lim_n\!\uparrow B_n$. It will be sufficient to show that for any $A_n, B_n \in \mathcal{F}_0$ we have

$$\lim_n\!\uparrow A_n \subset \lim_n\!\uparrow B_n \Longrightarrow \lim_n P_0[A_n] \leq \lim_n P_0[B_n]. \tag{10}$$

Notice that if $\lim_n\!\uparrow A_n \subset \lim_n\!\uparrow B_n$ then $A_n = A_n \cap \big(\lim_m\!\uparrow B_m\big) = \lim_m\!\uparrow (A_n \cap B_m)$. Therefore

$$P_0[A_n] = P_0\big[\lim_m\!\uparrow \underbrace{(A_n \cap B_m)}_{\in\mathcal{F}_0}\big]$$

$$= \lim_m \uparrow P_0[(A_n \cap B_m)], \quad \text{since } P_0 \text{ is a prob measure}$$
$$\le \lim_m \uparrow P_0[B_m].$$

Now taking limits of both sides in n gives $\lim_n \uparrow P_0[A_n] \le \lim_m \uparrow P_0\big[B_m\big]$ as was to be shown. Notice that (10) implies increasingness of $P^\uparrow$. In particular

$$A^\uparrow, B^\uparrow \in \mathcal{F}^\uparrow \text{ and } A^\uparrow \subset B^\uparrow \Longrightarrow P^\uparrow[A^\uparrow] \le P^\uparrow[B^\uparrow]. \tag{11}$$

(*Show* $(\mathcal{F}^\uparrow, P^\uparrow)$ *extends* $(\mathcal{F}_0, P_0)$) We need to show that $\mathcal{F}_0 \subset \mathcal{F}^\uparrow$ and $P^\uparrow = P_0$ on $\mathcal{F}_0$. Clearly $\mathcal{F}_0 \subset \mathcal{F}^\uparrow$ holds since any $A \in \mathcal{F}_0$ can be trivially written as $A = \lim_n \uparrow A$. The second statement follows since whenever $A \in \mathcal{F}_0$ we have that

$$P^\uparrow[A] := \lim_n \uparrow P_0[A] = P_0[\lim_n \uparrow A] = P_0[A] \tag{12}$$

where the second equality follows since we are assuming P_0 is a probability measure.

(*Show* P^* *is well defined*) Trivial.

(*Show* $(2^\Omega, P^*)$ *extends* $(\mathcal{F}^\uparrow, P^\uparrow)$) Trivially we have $\mathcal{F}^\uparrow \subset 2^\Omega$. Also notice that if $A \in \mathcal{F}^\uparrow$ then

$$P^\uparrow(A) \le \underbrace{\inf\{P^\uparrow(A^\uparrow) : A \subset A^\uparrow \in \mathcal{F}^\uparrow\}}_{=P^*[A]} \le P^\uparrow(A).$$

where the first inequality is given by (11) and the second inequality follows since $A \subset A \in \mathcal{F}^\uparrow$ is one of the covers in the infimum.

The proofs for $(\mathcal{F}^\downarrow, P^\downarrow)$ and $(2^\Omega, P_*)$ follow in a similar manner (after noticing $A_n \uparrow A \Longleftrightarrow A_n^c \downarrow A^c$).

□

Theorem 23 (5 facts about P^* and P_*). *For all sets $A, B, C, A_1, \ldots \in 2^\Omega$*

1. $P^*(A) + P_*(A^c) = 1$.
2. *If $A \subset B \subset C$ then $P_*(A) \le P_*(B) \le P^*(B) \le P^*(C)$.*
3. $P^*(A \cup B) \le P^*(A) + P^*(B) - P^*(A \cap B)$.
4. $P_*(A \cup B) \ge P_*(A) + P_*(B) - P_*(A \cap B)$
5. *If $A_n \uparrow A$ then $P^*(A_n) \uparrow P^*(A)$.*

Proof. These are a tedious and not very insightful so we will skip the proof in this class. □

Theorem 24 (Carathéodory extension theorem). *The probability measure P_0 on $\mathcal{F}_0$ has a unique extension to a probability measure P on $\sigma\langle\mathcal{F}_0\rangle =: \mathcal{F}$.*

Proof. Notice that the uniqueness follows from Theorem 19 since $\mathcal{F}_0$ is already a π-system. Therefore all we need to show is that $\bar{\mathcal{F}}$ is a σ-field containing $\mathcal{F}_0$ and $\bar{P}$ is a probability measure on $\bar{\mathcal{F}}$.

(*Show* $\mathcal{F}_0 \subset \bar{\mathcal{F}}$) In particular we need to show $A \in \mathcal{F}_0 \Longrightarrow P^*(A) = P_*(A)$. This follows directly by the fact that $(2^\Omega, P^*)$ and $(2^\Omega, P_*)$ are extensions of $(\mathcal{F}_0, P_0)$ by Theorem 22.

(*Show* $\bar{\mathcal{F}}$ *is a field*)

•($\Omega \in \bar{\mathcal{F}}$) Just showed $\mathcal{F}_0 \subset \bar{\mathcal{F}}$ and $\Omega \in \mathcal{F}_0$.

•($A \in \bar{\mathcal{F}} \Longrightarrow A^c \in \bar{\mathcal{F}}$) Suppose $A \in \bar{\mathcal{F}}$. Then

$$\begin{aligned} P^*(A^c) &= 1 - P_*(A), \quad \text{by Theorem 23.1} \\ &= 1 - P^*(A), \quad \text{since } A \in \bar{\mathcal{F}} \\ &= P_*(A^c), \quad \text{by Theorem 23.1.} \end{aligned} \tag{13}$$

Therefore $A^c \in \bar{\mathcal{F}}$.

•($A, B \in \bar{\mathcal{F}} \Longrightarrow A \cup B \in \bar{\mathcal{F}}$) Suppose $A, B \in \bar{\mathcal{F}}$. Then

$$\begin{aligned} P^*(A \cup B) &\le P^*(A) + P^*(B) - P^*(A \cap B), \quad \text{by Theorem 23.3} \\ &= P_*(A) + P_*(B) - P^*(A \cap B), \quad \text{since } A, B \in \bar{\mathcal{F}} \\ &\le P_*(A) + P_*(B) - P_*(A \cap B), \quad \text{by Theorem 23.2} \\ &\le P_*(A \cup B), \quad \text{by Theorem 23.4} \\ &\le P^*(A \cup B), \quad \text{by Theorem 23.2.} \end{aligned} \tag{14}$$

For one thing, this implies $P^*(A \cup B) = P_*(A \cup B)$ so that $A \cup B \in \bar{\mathcal{F}}$ as was to be shown.

(*Show* $\bar{\mathcal{F}}$ *is a monotone class*) Since $\bar{\mathcal{F}}$ is a field we simply show that $\bar{\mathcal{F}}$ is closed under monotonically increasing and decreasing limits. In particular, let $A_n \in \bar{\mathcal{F}}$ such that $A_n \uparrow A$. Then

$$\begin{aligned} P^*(A) &= \lim_n P^*(A_n), \quad \text{by Theorem 23.5} \\ &= \lim_n P_*(A_n), \quad \text{since } A_n \in \bar{\mathcal{F}} \\ &\le \lim_n P_*(A), \quad \text{by Theorem 23.2} \\ &= P_*(A) \\ &\le P^*(A), \quad \text{by Theorem 23.2} \end{aligned}$$

To show closure under decreasing limits just use the fact that $A_n \uparrow A \Longleftrightarrow A_n^c \downarrow A^c$ and equation (13).

(*Show* $\bar{P}$ *is a measure on* $\bar{\mathcal{F}}$) By Theorem 18 it will be sufficient to show that $\bar{P}$ is a FAP and $\bar{P}$ is continuous from below.

•($\bar{P}$ *is a FAP*) By extension facts $\bar{P}(\Omega) = P^*(\Omega) = P_0(\Omega) = 1$. (13) shows that $\bar{P}(\varnothing) = P^*(\varnothing) = 1 - P^*(\Omega) = 0$. Also, (14) establishes inclusion exclusion for P^* on $\bar{\mathcal{F}}$. Therefore whenever $A, B \in \bar{\mathcal{F}}$ and $A \cap B = \varnothing$ we get $\bar{P}(A \cup B) = \bar{P}(A) + \bar{P}(B)$. Therefor $\bar{P}$ is a FAP.

•($\bar{P}$ *is continuous from below*) Trivial from Theorem 23.5. □

Theorem 25. *$(\mathcal{F}, P)$ is an extension of both $(\mathcal{F}^\uparrow, P^\uparrow)$ and $(\mathcal{F}^\downarrow, P^\downarrow)$.*

Proof. Clearly both $\mathcal{F}^\uparrow \subset \mathcal{F}$ and $\mathcal{F}^\downarrow \subset \mathcal{F}$ by closure properties of $\mathcal{F}$ (in particular that any σ-field is also a monotone class). Let $A^\uparrow \in \mathcal{F}^\uparrow$. Then

$$\begin{aligned} P^\uparrow(A^\uparrow) &= P^*(A^\uparrow), \quad \text{since } P^* \text{ extends } P^\uparrow \text{ by Thm 22} \\ &= \bar{P}(A^\uparrow), \quad \text{since } A^\uparrow \in \bar{\mathcal{F}} \end{aligned}$$

$$= P(A^{\uparrow}), \quad \text{since } A^{\uparrow} \in \mathcal{F}.$$

Therefore P extends $P^{\uparrow}$. A similar proof establishes the desired result for $P^{\downarrow}$. $\square$

Section Assumption. *For the remainder of this section let P denote the probability measure on $\sigma\langle\mathcal{F}_0\rangle$ which is the unique extension of P_0 on $\mathcal{F}_0$. Also let $\mathcal{F} := \sigma\langle\mathcal{F}_0\rangle$.*

Theorem 26 (Easier formula for P^*). *For all $A \subset \Omega$*

1. $P^*(A) = \inf\{P(B) : A \subset B \in \mathcal{F}\}$;
2. $P_*(A) = \sup\{P(B) : A \supset B \in \mathcal{F}\}$.

Moreover, the above infimum and supremum are attained.

Proof.

$$\begin{aligned}
P^*(A) &:= \inf\{P^{\uparrow}(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\} \\
&= \inf\{P^*(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\}, \quad P^* \text{ extends } P^{\uparrow} \\
&\geq \inf\{\underbrace{P^*(B)}_{=P(B)} : A \subset B \in \mathcal{F}\}, \quad \text{inf over larger set} \\
&\geq P^*(A)
\end{aligned}$$

where the last inequality follows since $P^*(B) \geq P^*(A)$ (by Theorem 23.2). A similar proof establishes the result for P_*.

To see why the infimum is attained let $A \subset B_n \in \mathcal{F}$ such that $P(B_n) \to P^*(A)$. Now

$$P\Big(\bigcap_{n=1}^{\infty} B_n\Big) = \lim_N\downarrow P\Big(\bigcap_{n=1}^{N} B_n\Big) \leq \lim_N P(B_N) = P^*(A). \quad (15)$$

Therefore

$$\begin{aligned}
P\Big(\bigcap_{n=1}^{\infty} B_n\Big) &\leq P^*(A), \quad \text{by (15)} \\
&= \inf\{P(B) : A \subset B \in \mathcal{F}\} \\
&\leq P\Big(\bigcap_{n=1}^{\infty} B_n\Big)
\end{aligned}$$

where the last inequality follows from the fact that $A \subset \bigcap_{n=1}^{\infty} B_n \in \mathcal{F}$. Therefore the infimum is attained as was to be shown. A similar proof is used for the supremum. $\square$

Although the above infimum and supremum are attained in Theorem notice that the following infimum and supremum are **not** necessarily attained:

$$\begin{aligned}
P^*(A) &:= \inf\{P^{\uparrow}(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\} \\
P_*(A) &:= \sup\{P^{\downarrow}(A^{\downarrow}) : A \supset A^{\downarrow} \in \mathcal{F}^{\downarrow}\}.
\end{aligned}$$

The following theorem is as close as we can get working with $\mathcal{F}^{\uparrow}$ and $\mathcal{F}^{\downarrow}$.

Theorem 27 (Approximating P with $\mathcal{F}^{\uparrow}$). *For all $A \in \mathcal{F}$ there exists $\mathcal{F}^{\downarrow}$-sets $A_n^{\downarrow}$ and $\mathcal{F}^{\uparrow}$-sets $A_n^{\uparrow}$ such that*

- $\bigcup_{n=1}^{\infty} A_n^{\downarrow} \subset A \subset \bigcap_{n=1}^{\infty} A_n^{\uparrow}$;
- $P\Big(\bigcup_{n=1}^{\infty} A_n^{\downarrow}\Big) = P(A) = P\Big(\bigcap_{n=1}^{\infty} A_n^{\uparrow}\Big)$.

Proof. Let $A \subset A_n^{\uparrow} \in \mathcal{F}^{\uparrow}$ such that

$$\begin{aligned}
\lim_n P(A_n^{\uparrow}) &= \inf\{P^{\uparrow}(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\} \\
&= \inf\{P(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\}.
\end{aligned}$$

By a similar proof as in Theorem 3.1 we get $A \subset \bigcap_{n=1}^{\infty} A_n^{\uparrow} \in \mathcal{F}$ and

$$P\Big(\bigcap_{n=1}^{\infty} A_n^{\uparrow}\Big) \leq \underbrace{\inf\{P(A^{\uparrow}) : A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\}}_{= P(A) \text{ when } A \in \mathcal{F}} \leq P\Big(\bigcap_{n=1}^{\infty} A_n^{\uparrow}\Big).$$

The proof for $\mathcal{F}^{\downarrow}$ is similar. $\square$

Theorem 28 (Approximating P with $\mathcal{F}_0$). *For all $A \in \mathcal{F}$ and all $\epsilon > 0$ there exists $A^o \in \mathcal{F}_0$ such that*

- $P(A \triangle A^o) \leq \epsilon$.

Proof. If $A \in \mathcal{F}$ then $P(A) = P^*(A) = \inf\{P(A^{\uparrow}) \colon A \subset A^{\uparrow} \in \mathcal{F}^{\uparrow}\}$, since P^* extends P. Therefore one can find $A^{\uparrow} \in \mathcal{F}^{\uparrow}$ such that $A \subset A^{\uparrow}$ and

$$P(A^{\uparrow} - A) = P(A^{\uparrow}) - P(A) \leq \epsilon/2. \quad (16)$$

Note that $P(A^{\uparrow}) = P(\lim_n\uparrow A_n^o) = \lim_n\uparrow P(A_n^o)$ where $A_n^o \in \mathcal{F}_0$ and $A_n^o \subset A^{\uparrow}$. Therefore we can find A_n^o such that

$$P(A^{\uparrow} - A_n^o) = P(A^{\uparrow}) - P(A_n^o) \leq \epsilon/2. \quad (17)$$

Now

$$\begin{aligned}
P(A \triangle A_n^o) &\leq P(A \cap (A_n^o)^c) + P(A^c \cap A_n^o) \\
&\leq P(A^{\uparrow} \cap (A_n^o)^c) + P(A^c \cap A^{\uparrow}) \\
&= P(A^{\uparrow} - A_n^o) + P(A^{\uparrow} - A) \\
&\leq \epsilon, \text{ by (16) and (17)}
\end{aligned}$$

$\square$

Theorem 29 (Use P^* to find P-neg sets). *Let $A \subset \Omega$*

$$\begin{aligned}
A \textit{ is } P\textit{-negligible} &\iff P^*(A) = 0 \quad (18) \\
&\implies A \in \bar{\mathcal{F}}. \quad (19)
\end{aligned}$$

Proof. (*Show (19)*) This follows since $0 \leq P_*(A) \leq P^*(A)$ for all $A \subset \Omega$ by Theorem 23.2.

(*Show $\Longrightarrow$ of (18)*) This follows since

$$P^*(A) = \inf\{\underbrace{P(B)}_{\text{one of these is } 0} : A \subset B \in \mathcal{F}\}. \quad (20)$$

(*Show $\Longleftarrow$ of (18)*) This follows since the infimum in (20) is attained so that there exists some $B \in \mathcal{F}$ such that $A \subset B$ and

$$\underbrace{0 = P^*(A)}_{\text{by assumption}} = P(B).$$

$\square$

The nice thing about the above theorem is that you can show both $\bar{P}(A) = 0$ and $A \in \bar{\mathcal{F}}$ just by establishing $P^*(A) = 0$, which you can technically analyze without knowing A is in $\mathcal{F}$ or $\bar{\mathcal{F}}$.

Theorem 30 **(The structure of** $(\Omega, \bar{\mathcal{F}}, \bar{P})$**).** *Let $\mathcal{N}_P \subset 2^\Omega$ denote the P-negligible sets. Then*

- $\bar{\mathcal{F}} = \sigma\langle \mathcal{F}, \mathcal{N}_P\rangle = \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}_P\}$;
- $\bar{P}[F \cup N] = P[F]$ *for all* $F \in \mathcal{F}$ *and* $N \in \mathcal{N}_P$.

Proof. Start by letting

$$\widetilde{\mathcal{F}} := \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}_P\}.$$

(*Show* $\bar{\mathcal{F}} \subset \widetilde{\mathcal{F}}$) Let $C \in \bar{\mathcal{F}}$ and we try to write C in the form $F \cup N$ where $F \in \mathcal{F}$ and N is P-negligible. Since $C \in \bar{\mathcal{F}}$ we have that

$$P_*[C] = \bar{P}[C] = P^*[C].$$

Since the infimum and supermum in Theorem 3.1 are attained, there exists $C^* \in \mathcal{F}$ and $C_* \in \mathcal{F}$ such that $C_* \subset C \subset C^*$ and

$$P[C_*] = \bar{P}[C] = P[C^*].$$

Now we have that

$$C = C_* \cup (C - C_*).$$

If we can show that $(C - C_*)$ is P-negligible we are done (in particular $C \in \widetilde{\mathcal{F}}$). To see why notice that $C - C_* \subset C^* - C_* \in \mathcal{F}$ which then implies

$$P[C - C_*] \le P[C^* - C_*] = P[C^*] - P[C_*] = 0.$$

(*Show* $\widetilde{\mathcal{F}} \subset \bar{\mathcal{F}}$ *and* $\bar{P}[F \cup N] = P[F]$) Let $F \cup N \in \widetilde{\mathcal{F}}$. It will be sufficient to show $P^*[F \cup N] = P_*[F \cup N] = P[F]$. To see why

$$\begin{aligned}
P[F] &= P_*[F], \quad \text{since } F \in \mathcal{F}\\
&\le P_*[F \cup N], \quad \text{by Theorem 23.2}\\
&\le P^*[F \cup N], \quad \text{by Theorem 23.2}\\
&\le P^*[F \cup B], \quad \text{where } N \subset B \in \mathcal{F},\ P[B] = 0\\
&\le P^*[F] + \underbrace{P^*[B]}_{=P[B]=0} - \underbrace{P^*[F \cap B]}_{\le P^*[B]=0}, \quad \text{by Theorem 23.3}\\
&= P[F].
\end{aligned}$$

(*Show* $\widetilde{\mathcal{F}} \subset \sigma\langle \mathcal{F}, \mathcal{N}_P\rangle$) This is obvious since $F \cup N \in \sigma\langle \mathcal{F}, \mathcal{N}_P\rangle$ for any $F \in \mathcal{F}$ and $N \in \mathcal{N}_P$.

(*Show* $\sigma\langle \mathcal{F}, \mathcal{N}_P\rangle \subset \widetilde{\mathcal{F}}$) This follows by *good sets*. Indeed $\widetilde{\mathcal{F}}$ is a σ-field since it equals the σ-field $\bar{\mathcal{F}}$. Also clearly $\mathcal{F} \subseteq \bar{\mathcal{F}} = \widetilde{\mathcal{F}}$. To finish we note that $\mathcal{N}_P \subset \widetilde{\mathcal{F}}$ since $N = \varnothing \cup N \in \widetilde{\mathcal{F}}$ for any $N \in \mathcal{N}_P$. ☐

Theorem 31 **(Regularity).** *Let μ be any measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which assigns finite measure to bounded sets in $\mathcal{B}(\mathbb{R}^d)$. For any $B \in \mathcal{B}(\mathbb{R}^d)$ and $\epsilon > 0$ there exists a closed set C and an open set O such that $C \subset B \subset O$ and*

$$\mu(O - C) < \epsilon.$$

Corollary 1. *Let μ be any measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which assigns finite measure to bounded sets in $\mathcal{B}(\mathbb{R}^d)$. Then*

$$\begin{aligned}
\mu(B) &= \sup\{\mu(C) : C \subset B,\ C \text{ closed}\}\\
&= \inf\{\mu(O) : B \subset O,\ O \text{ open}\}
\end{aligned}$$

Exercise 17. *Prove Theorem 31 for $d = 1$*

Exercise 18. *(a) Prove Corollary 1 for $d = 1$. (b) Give an example of a σ-finite measure μ on $\mathcal{B}(\mathbb{R})$ and a Borel set B such that*

$$\mu(B - C) = \infty = \mu(O - B)$$

for every closed subset C of B and every open superset O of B.

Exercise 19. *Suppose $\mathcal{F}_0$ is a field, μ is a measure on $\sigma\langle \mathcal{F}_0\rangle$ and μ is σ-finite on $\mathcal{F}_0$.*

1. *Suppose $B \in \sigma\langle \mathcal{F}_0\rangle$ and $\epsilon > 0$. Show that there exists a disjoint sequence of $\mathcal{F}_0$-sets $A_1, A_2, \ldots$ such that $B \subset \cup_{n=1}^\infty A_n$ and $\mu\big(\cup_{n=1}^\infty A_n - B\big) \le \epsilon$.*
2. *Suppose $B \in \sigma\langle \mathcal{F}_0\rangle$, $\mu(B) < \infty$ and $\epsilon > 0$. Show there exists an $\mathcal{F}_0$-set A such that $\mu(A \triangle B) \le \epsilon$.*
3. *Show by example that the conclusion to 2 may fail if B has infinite measure.*

Exercise 20. *(a) Show that $(\Omega, \bar{\mathcal{F}}, \bar{P})$ is the smallest complete extension of $(\Omega, \mathcal{F}, P)$—that is, if $(\Omega, \mathcal{F}', P')$ is probability space which is a complete extension of $(\Omega, \mathcal{F}, P)$, then $(\Omega, \mathcal{F}', P')$ is also a complete extension of $(\Omega, \bar{\mathcal{F}}, \bar{P})$. (b) Show by example that $(\Omega, \bar{\mathcal{F}}, \bar{P})$ can have infinitely many different complete extensions (Hint: use a sample space consisting of two points).*

3.2 Application: Lebesque Measure

Theorem 32. *The mapping $P : \mathcal{B}_0((0, 1]) \to [0, 1]$ defined in Definition 2 is a probability measure.*

Proof. We will use Theorem 18 and show that P is continuous from above at $\varnothing$. Let $A_n \downarrow \varnothing$ where $A_n \in \mathcal{B}_0((0, 1])$ (in particular we have that $\bigcap_{k=1}^\infty A_k = \varnothing$). We show $P[A_n] \downarrow 0$.

Notice first that for all $n \ge N$ we have

$$P[A_n] \le P[A_N] = P\big[\textstyle\bigcap_{k=1}^N A_k\big]$$

since the A_k's are decreasing. It will then be sufficient to show that for all $\epsilon > 0$ there exists N_ϵ such that

$$P\big[\textstyle\bigcap_{k=1}^{N_\epsilon} A_k\big] \le \epsilon. \tag{21}$$

The following argument doesn't quite work but it will motivate the solution. Let $\epsilon > 0$ and for each A_k find a sequence of closed sets F_k such that $F_k \subset A_k$ and $P[A_k - F_k] \le \epsilon/2^k$ (If A_k has the form $\bigcup_i (a_i, b_i]$ take $F_k := \bigcup_i [a_i + \tau, b_i]$ for small enough τ). Since

$\bigcap_{k=1}^{\infty} A_k = \varnothing$ one has that $\bigcap_{k=1}^{\infty} F_k = \varnothing$. By a compactness argument[1] there exists an N_ϵ such that $\bigcap_{k=1}^{N_\epsilon} F_k = \varnothing$. Therefore

$$P\big[\bigcap_{k=1}^{N_\epsilon} A_k\big] - \underbrace{P\big[\underbrace{\bigcap_{k=1}^{N_\epsilon} F_k}_{=\varnothing}\big]}_{=0} \le \sum_{k=1}^{N_\epsilon} P\big[A_k - F_k\big] \le \sum_{k=1}^{N_\epsilon} \tfrac{1}{2^k} \le \epsilon.$$

This would establish (21) if it weren't for the problem that P isn't defined on the F_k's.

It is clear how to fix this. For each A_k find closed sets F_k and $\mathcal{B}_0((0,1])$-sets A_k^o such that $A_k \supset F_k \supset A_k^o$ and $P[A_k - A_k^o] \le \epsilon/2^k$. For each $\epsilon > 0$ we still have the property that there exists N_ϵ such that $\bigcap_{k=1}^{N_\epsilon} F_k = \varnothing$ which implies $\bigcap_{k=1}^{N_\epsilon} A_k^o = \varnothing$ and now

$$P\big[\bigcap_{k=1}^{N_\epsilon} A_k\big] - \underbrace{P\big[\underbrace{\bigcap_{k=1}^{N_\epsilon} A_k^o}_{=\varnothing}\big]}_{=0} \le \sum_{k=1}^{N_\epsilon} P\big[A_k - A_k^o\big] \le \sum_{k=1}^{N_\epsilon} \tfrac{1}{2^k} \le \epsilon.$$

□

Theorem 33 (Application to Borel's normal numbers). *Let $P : \mathcal{B}_0((0,1]) \to [0,1]$ be as in Definition 2. Then*

1. *P has a unique extension to a probably measure on $\mathcal{B}((0,1])$ (still denoted P for the rest of this theorem);*

2. *P is the only measure on $\mathcal{B}((0,1])$ which satisfies $P[(0,x]] = x$ for all $x \in (0,1]$;*

3. *$N \in \mathcal{B}((0,1])$ and $P[N] = 1$ where N is the set of normal numbers in $(0,1]$;*

4. *$\mathcal{B}_0((0,1]) \subsetneq \mathcal{B}((0,1]) \subsetneq \overline{\mathcal{B}((0,1])} \subsetneq 2^\Omega$. Sets in $\mathcal{B}((0,1])$ are called* **Borel measurable**. *Sets in $\overline{\mathcal{B}((0,1])}$ are called* **Lebesgue measurable**.

Proof. (*Show item 1*) First note the Carathéodory Extension Theorem along with Theorem 32 shows there exists a unique extension $P : \mathcal{B}_0((0,1]) \to [0,1]$ to $P : \mathcal{B}((0,1]) \to [0,1]$ since $\mathcal{B}((0,1]) = \sigma\langle \mathcal{B}_0((0,1])\rangle$.

(*Show item 2*) This follows from the uniqueness theorem for measures since $\mathcal{B}((0,1]) = \sigma\langle (0,x] \colon x \in (0,1]\rangle$ and $\{(0,x] \colon x \in (0,1]\}$ is a π-system.

(*Show item 3*) Notice first that N and N^c are both in $\mathcal{B}((0,1])$. Now, in Theorem 3 we showed that N^c is negligible. In particular, for any $\epsilon > 0$, there exists $B_n \in B_0((0,1])$ such that $N^c \subset \bigcup_{n=1}^{\infty} B_n$ where $\sum_{n=1}^{\infty} P[B_n] \le \epsilon$. Since $\bigcup_{n=1}^{\infty} B_n \in \mathcal{B}((0,1])$ we have

$$P\Big(\bigcup_{n=1}^{\infty} B_n\Big) \le \sum_{n=1}^{\infty} P[B_n] \le \epsilon.$$

Therefore

$$P(N^c) = \inf\{P(B) \colon N^c \subset B \in \mathcal{F}\} \le \epsilon$$

for all ϵ. Therefore $P(N^c) = 0$ and $P(N) = 1$.

(*Show item 4*) We've already mentioned that $N \in \mathcal{B}((0,1])$ but $N \notin \mathcal{B}_0((0,1])$. Exercise 15 of page 15 in Chung ("A Course in Probability Theory") shows that $\mathcal{B}((0,1]) \subsetneq \overline{\mathcal{B}((0,1])}$. Billingsley ("Probability and Measure") page 46 shows that it is impossible to extend P to a probability measure on 2^Ω which establishes that $\overline{\mathcal{B}((0,1])} \subsetneq 2^\Omega$.

□

For any $\boldsymbol{i} = (i_1, \ldots, i_d) \in \mathbb{Z}^d$ let $(\boldsymbol{i}, \boldsymbol{i}+1]$ be the unit cube in $\mathbb{R}^d$ translated up by $\boldsymbol{i}$ so that

$$(\boldsymbol{i}, \boldsymbol{i}+1] \equiv (i_1, i_1+1] \times \cdots \times (i_d, i_d+1].$$

Notice that these sets give a checker board decomposition, $\mathbb{R}^d = \bigcup_{\boldsymbol{i} \in \mathbb{Z}^d} (\boldsymbol{i}, \boldsymbol{i}+1]$, so that $\mathbb{R}^d$ is expressed as a countable disjoint union of the translated unit cubes. Let $\mathcal{B}_0^{(\boldsymbol{i},\boldsymbol{i}+1]}$ denote the field of finite disjoint unions of rectangles in $(\boldsymbol{i}, \boldsymbol{i}+1]$ and let $\mathcal{B}((\boldsymbol{i}, \boldsymbol{i}+1]) \equiv \sigma\langle \mathcal{B}_0^{(\boldsymbol{i},\boldsymbol{i}+1]}\rangle$ denote the Borel σ-field of $(\boldsymbol{i}, \boldsymbol{i}+1]$. Finally let $P_{\boldsymbol{i}}$ denote the unique uniform probability measure on $\mathcal{B}((\boldsymbol{i}, \boldsymbol{i}+1])$ which assigns Euclidean volume to the rectangles in $(\boldsymbol{i}, \boldsymbol{i}+1]$, i.e.

$$P_{\boldsymbol{i}}\big((a_1, b_1] \times \cdots \times (a_d, b_d]\big) = \prod_{k=1}^{d} (b_k - a_k)$$

whenever $(a_1, b_1] \times \cdots \times (a_d, b_d] \subset (\boldsymbol{i}, \boldsymbol{i}+1]$. The construction of $P_{\boldsymbol{i}}$ is done in exactly the same way as the uniform probability measure was constructed on $(0,1]$ in the beginning of the class. Lets recall how this is done. One first shows that for any $A \in B_0^{(\boldsymbol{i},\boldsymbol{i}+1]}$ one can define $P_{\boldsymbol{i}}(A)$ to be the sum of the disjoint rectangle volumes which make up A (this is not trivial since there are different decompositions of A into disjoint rectangles, but one can use a result similar to Theorem 1.3 of Billingsley to prove that $P_{\boldsymbol{i}}$ is well defined). Secondly, one shows that $P_{\boldsymbol{i}}$ is a probability measure on $((\boldsymbol{i}, \boldsymbol{i}+1], \mathcal{B}_0^{(\boldsymbol{i},\boldsymbol{i}+1]})$. The hard part of this step is to show the countable additivity. For $(0,1]$ we used the equivalent condition that $P_{\boldsymbol{i}}$ is continuous from above at $\varnothing$. This argument carries over to $((\boldsymbol{i}, \boldsymbol{i}+1], \mathcal{B}_0^{(\boldsymbol{i},\boldsymbol{i}+1]}, P_{\boldsymbol{i}})$. Finally one invokes the Carathéodory Extension theorem to get a uniform probability measure $((\boldsymbol{i}, \boldsymbol{i}+1], \mathcal{B}((\boldsymbol{i}, \boldsymbol{i}+1]), P_{\boldsymbol{i}})$ (uniqueness follows by the fact that rectangles, including the empty ones, form a π-system).

Now, using the uniform probability measures $((\boldsymbol{i}, \boldsymbol{i}+1], \mathcal{B}((\boldsymbol{i}, \boldsymbol{i}+1]), P_{\boldsymbol{i}})$ we can define Lebesque measure $\mathcal{L}^d$ on sets $A \in \mathcal{B}((\boldsymbol{i}, \boldsymbol{i}+$

[1] For, if not, then there exists $x_n \in \bigcap_{k=1}^{n} F_k$ for each n. Notice

$$\bigcap_{k=1}^{n} F_k \subset \bigcap_{k=1}^{m} F_k \text{ when } m \le n \tag{22}$$

For one thing, equation (22) implies that $x_n \in F_1$. Therefore by compactness there exists a sub-sequential limit $x = \lim_k x_{n_k} \in F_1$. Again by (22) and the assumption $x_{n_k} \in \bigcap_{k=1}^{n_k} F_k$ one has that for sufficiently large k all x_{n_k} are eventually within F_m. Therefore $x \in F_m$ for each m. This contradicts the assumption $\cap_{k=1}^{\infty} F_k = \varnothing$.

1]) by stitching these $P_{\boldsymbol{i}}$ together as follows

$$\mathcal{L}^d(A) := \sum_{\boldsymbol{i}\in\mathbb{Z}^d} P_{\boldsymbol{i}}\big((\boldsymbol{i},\boldsymbol{i}+1]\cap A\big). \tag{23}$$

Notice that each $(\boldsymbol{i},\boldsymbol{i}+1]\cap A$ is in the Borel σ-field $\mathcal{B}((\boldsymbol{i},\boldsymbol{i}+1])$ by Claim 10 so that $P_{\boldsymbol{i}}\big((\boldsymbol{i},\boldsymbol{i}+1]\cap A\big)$ is defined. Lets see that $\mathcal{L}^d$ is indeed a measure on $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$.

Theorem 34. *$\mathcal{L}^d$ is a measure on $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$.*

Proof. We show the following three axioms (i), (ii) and (iii):

(i) $\mathcal{L}^d(A)\in[0,\infty]$: Trivial.

(ii) $\mathcal{L}^d(\varnothing)=0$: This is also easy since $P_{\boldsymbol{i}}\big((\boldsymbol{i},\boldsymbol{i}+1]\cap\varnothing\big)=0$.

(iii) Countable additivity: Suppose $A_1,A_2,\ldots\in\mathcal{B}(\mathbb{R}^d)$ are disjoint. Then

$$\begin{aligned}
\mathcal{L}^d\Big(\bigcup_{k=1}^\infty A_k\Big) &= \sum_{\boldsymbol{i}\in\mathbb{Z}^d} P_{\boldsymbol{i}}\Big((\boldsymbol{i},\boldsymbol{i}+1]\cap\bigcup_{k=1}^\infty A_k\Big)\\
&= \sum_{\boldsymbol{i}\in\mathbb{Z}^d} P_{\boldsymbol{i}}\Big(\bigcup_{k=1}^\infty(\boldsymbol{i},\boldsymbol{i}+1]\cap A_k\Big)\\
&= \sum_{\boldsymbol{i}\in\mathbb{Z}^d}\sum_{k=1}^\infty P_{\boldsymbol{i}}\Big((\boldsymbol{i},\boldsymbol{i}+1]\cap A_k\Big) \qquad (24)\\
&= \sum_{k=1}^\infty\sum_{\boldsymbol{i}\in\mathbb{Z}^d} P_{\boldsymbol{i}}\Big((\boldsymbol{i},\boldsymbol{i}+1]\cap A_k\Big) \qquad (25)\\
&= \sum_{k=1}^\infty \mathcal{L}^d\big(A_k\big)
\end{aligned}$$

where (24) follows since $P_{\boldsymbol{i}}$ is countably additive and the $(\boldsymbol{i},\boldsymbol{i}+1]\cap A_k$'s are disjoint; and (25) follows from general results about positive iterated sums.

□

Theorem 35. *$\mathcal{L}^d$ is the only measure on $(\mathbb{R}^d,\mathcal{B}((0,1]^d))$ which assigns standard Euclidean volume to the finite rectangles as follows*

$$\mathcal{L}^d\big((a_1,b_1]\times\cdots\times(a_d,b_d]\big)=\prod_{k=1}^d(b_k-a_k) \tag{26}$$

for $-\infty<a_k<b_k<\infty$.

Proof. Define $\mathcal{P}$ to be the π-system composed of the finite rectangles $\{(a_1,b_1]\times\cdots\times(a_d,b_d]:-\infty<a_k<b_k<\infty\}$ and the empty set $\varnothing$. One can easily establish that $\mathcal{B}(\mathbb{R}^d)=\sigma\langle\mathcal{P}\rangle$. Also notice that $\mathcal{L}^d$ is σ-finite on $\mathcal{P}$ since $\mathcal{L}^d\big((\boldsymbol{i},\boldsymbol{i}+1]\big)=1$, $\mathbb{R}^d=\cup_{\boldsymbol{i}\in\mathbb{Z}^d}(\boldsymbol{i},\boldsymbol{i}+1]$ and each $(\boldsymbol{i},\boldsymbol{i}+1]\in\mathcal{P}$. Therefore Theorem 20 establishes the following claim □

Theorem 36. *For any $A\in\mathcal{B}^{\mathbb{R}^d}$ and $x\in\mathbb{R}^d$, the set $A+x:=\{a+x:a\in A\}$ is in $\mathcal{B}(\mathbb{R}^d)$ and*

$$\mathcal{L}^d(A+x)=\mathcal{L}^d(A) \tag{27}$$

Proof. To show $A+x\in\mathcal{B}(\mathbb{R}^d)$ use the good sets principle. Fix $x\in\mathbb{R}^d$ and set $\mathcal{G}_x:=\{A\in\mathcal{B}(\mathbb{R}^d):A+x\in\mathcal{B}(\mathbb{R}^d)\}$. It is easy to see that $\mathcal{G}_x$ is a σ-field since complementation and union is preserved under translation by x. For example,

$$\begin{aligned}
A\in\mathcal{G}_x &\Rightarrow A\in\mathcal{B}(\mathbb{R}^d)\text{ and }A+x\in\mathcal{B}(\mathbb{R}^d)\\
&\Rightarrow A^c\in\mathcal{B}(\mathbb{R}^d)\text{ and }(A+x)^c\in\mathcal{B}(\mathbb{R}^d)\\
&\Rightarrow A^c\in\mathcal{B}(\mathbb{R}^d)\text{ and }A^c+x\in\mathcal{B}(\mathbb{R}^d)\\
&\Rightarrow A^c\in\mathcal{G}_x.
\end{aligned}$$

The other axioms are established in a similar fashion. Moreover, clearly all the finite rectangles are in $\mathcal{G}_x$. Therefore good sets implies $\mathcal{B}(\mathbb{R}^d)\subset\mathcal{G}_x$ which implies $A\in\mathcal{B}(\mathbb{R}^d)\to A+x\in\mathcal{B}(\mathbb{R}^d)$, as was to be shown.

Now to show (27) one can simply use the same arguments used in the Theorem 35 on the uniqueness of $\mathcal{L}^d$. In particular, fix x and define $\mu_x(A):=\mathcal{L}^d(A+x)$. It is easy to show that μ_x is a measure on $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$. Moreover, since the volume of any rectangle in $\mathbb{R}^d$ is invariant under translation by x, the measures μ_x and $\mathcal{L}^d$ both agree on the π-system of finite, possibly empty, rectangles in $\mathbb{R}^d$. Since they are also both σ-finite on these rectangles one must have, by Theorem 35, $\mathcal{L}^d(A)=\mu_x(A):=\mathcal{L}^d(A+x)$ for all $A\in\mathcal{B}(\mathbb{R}^d)$, as was to be shown. □

Theorem 37. *If $T:\mathbb{R}^d\to\mathbb{R}^d$ is linear and nonsingular, then $A\in\mathcal{B}(\mathbb{R}^d)$ implies that $TA:=\{T(a):a\in A\}\in\mathcal{B}(\mathbb{R}^d)$ and*

$$\mathcal{L}^d(TA):=|\det T|\mathcal{L}^d(A).$$

Theorem 38. *Let $(\Omega,\mathcal{F},\mu)$ be a σ-finite measure space. Then $\mathcal{F}$ cannot contain an uncountable, disjoint collection of sets of positive μ-measure*

Proof. Let $\{B_i:i\in\mathcal{I}\}$ a disjoint collection of sets of such that $\mu(B_i)>0$ for each $i\in\mathcal{I}$. We show $\mathcal{I}$ must be countable.

Since μ is σ-finite there exists $A_1,A_2,\ldots\in\mathcal{F}$ such that $\mu(A_k)<\infty$ and $\Omega=\cup_kA_k$. We show the following three facts.

- $\{i\in\mathcal{I}:\mu(A_k\cap B_i)>\epsilon\}$ is finite for all k: Let $\epsilon>0$ and suppose by contradiction one can find a countably infinite set $\mathcal{I}_c\subset\mathcal{I}$ such that $\mu(A_k\cap B_i)>\epsilon$ for all $i\in\mathcal{I}_c$ and for this set of indices one has

  $$\mu(A_k)\ge\mu(A_k\cap(\cup_{i\in\mathcal{I}_c}B_i))=\sum_{i\in\mathcal{I}_c}\mu(A_k\cap B_i)>\sum_{i\in\mathcal{I}_c}\epsilon=\infty$$

  which gives a contradiction.

- $\{i\in\mathcal{I}:\mu(A_k\cap B_i)>0\}$ is countable for all k: This follows from the identity

  $$\{i\in\mathcal{I}:\mu(A_k\cap B_i)>0\}=\bigcup_{\text{rational }\epsilon}\underbrace{\{i\in\mathcal{I}:\mu(A\cap B_i)>\epsilon\}}_{\text{finite by (i)}}.$$

- $\mathcal{I}=\bigcup_k\{i\in\mathcal{I}:\mu(A_k\cap B_i)>0\}$: To show $\mathcal{I}\cup\bigcup_k\{i\in\mathcal{I}:\mu(A_k\cap B_i)>0\}$ notice that if $i\in\mathcal{I}$ then $\mu(B_i)>0$. Now $\Omega=\cup_kA_k$ so there must exist a k such that $\mu(A_k\cap B_i)>0$. Therefore $i\in\bigcup_k\{i\in\mathcal{I}:\mu(A_k\cap B_i)>0\}$. The other inclusion is obvious.

To finish the proof simply notice that the last two bullets imply $\mathcal{I}$ is countable. $\square$

Corollary 2. *If $k < d$ then $\mathcal{L}^d(A) = 0$ for any k-dimensional hyperplane $A \subset \mathbb{R}^d$ where $k < d$.*

Proof. Let A be a k-dimensional hyperplane where $k < d$. Let x be a point in $\mathbb{R}^d$ which is not contained in A. Then $\{A + xt : t \in \mathbb{R}\}$ is an uncountable class of disjoint subsets of $\mathcal{B}(\mathbb{R}^d)$. Since $\mathcal{L}^d$ is translation invariance $\mathcal{L}^d(A) = \mathcal{L}^d(A + xt)$ for each $t \in \mathbb{R}$. Now by Theorem 38, $\mathcal{L}^d(A) = 0$, for otherwise there would exists a uncountable, disjoint collection of sets of positive $\mathcal{L}^d$-measure. $\square$

Definition 25 (**Borel versus Lebesque measurable sets**). *Let $\big(\Omega, \overline{\mathcal{B}(\mathbb{R}^d)}, \overline{\mathcal{L}^d}\big)$ be the completion of $(\Omega, \mathcal{B}(\mathbb{R}^d), \mathcal{L}^d)$. If $A \in \mathcal{B}(\mathbb{R}^d)$ then A is said to be Borel measurable. If $A \in \overline{\mathcal{B}(\mathbb{R}^d)}$ then A is said to be Lebesque measurable.*

Theorem 39 (**This is why we need σ-fields**).

- $\mathcal{B}(\mathbb{R}) \subsetneq \overline{\mathcal{B}(\mathbb{R})} \subsetneq 2^{\mathbb{R}}$.
- *It is impossible to put a measure on $2^{\mathbb{R}}$ which is translation invariant and which assigns normal length to finite intervals. Put another way—there is no Lebesque measure on all of $2^{\mathbb{R}}$. Or another way—it is impossible to consistently assign a length to all subsets of $\mathbb{R}$.*

4 Independence for classes of events

Definition 26 (**Probability space).** *If Ω is a sample space, $\mathcal{F}$ is a σ-field on Ω and P is a probability measure on $\mathcal{F}$, the triple $(\Omega, \mathcal{F}, P)$ is called a probability space.*

Section Assumption. *Throughout this section let $(\Omega, \mathcal{F}, P)$ denote a probability space and $\mathcal{K}$ be an arbitrary index set.*

Definition 27 (**Independent events).** *A collection of $\mathcal{F}$-sets $\{A_k\}_{k\in\mathcal{K}}$ are said to be independent if for every finite index set $\mathcal{H} \subset \mathcal{K}$ the following identity holds:*

$$P\Big(\bigcap_{h\in\mathcal{H}} A_h\Big) = \prod_{h\in\mathcal{H}} P(A_h).$$

Definition 28 (**Independent classes).** *Let $\mathscr{A}_k$ be a collection of $\mathcal{F}$-sets for each $k \in \mathcal{K}$ (i.e. $\mathscr{A}_k \subset \mathcal{F}$). Then $\{\mathscr{A}_k\}_{k\in\mathcal{K}}$ are independent classes if for each choice $A_k \in \mathscr{A}_k$ the events $\{A_k\}_{k\in\mathcal{K}}$ are independent.*

Theorem 40.

1. **(Subclasses).** *If $\mathscr{A}_k \subset \mathscr{B}_k \subset \mathcal{F}$ for all $k \in \mathcal{K}$ and $\{\mathscr{B}_k\}_{k\in\mathcal{K}}$ are independent classes then $\{\mathscr{A}_k\}_{k\in\mathcal{K}}$ are independent classes.*
2. **(Augmentation).** *$\{\mathscr{A}_k\}_{k\in\mathcal{K}}$ are independent classes if and only if $\{\mathscr{A}_k \cup \{\Omega\}\}_{k\in\mathcal{K}}$ are independent classes.*
3. **(Simplified product).** *If $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are collections of $\mathcal{F}$-sets and $\Omega \in \mathscr{A}_k$ for each k, then $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are independent classes if and only if*

$$P\Big(\bigcap_{k=1}^n A_k\Big) = \prod_{k=1}^n P(A_k).$$

for each choice $A_k \in \mathscr{A}_k$.

Theorem 41 (**π-generators are enough).** *Suppose $\{\mathscr{A}_k\}_{k\in\mathcal{K}}$ are independent classes of $\mathcal{F}$-sets such that each $\mathscr{A}_k$ is also a π-system. Then $\{\sigma\langle\mathscr{A}_k\rangle\}_{k\in\mathcal{K}}$ are independent classes.*

Theorem 42 (**ANOVA).** *Let $\mathscr{A}_1, \mathscr{A}_2, \ldots$ and $\mathscr{B}_1, \mathscr{B}_2, \ldots$ be classes of $\mathcal{F}$-sets which are π-systems. Then $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{B}_1, \mathscr{B}_2, \ldots$ are all independent if and only if the following three statements hold:*

1. *$\mathscr{A}_1, \mathscr{A}_2, \ldots$ are independent;*
2. *$\mathscr{B}_1, \mathscr{B}_2, \ldots$ are independent;*
3. *$\sigma\langle\mathscr{A}_1, \mathscr{A}_2, \ldots\rangle$ is independent of $\sigma\langle\mathscr{B}_1, \mathscr{B}_2, \ldots\rangle$.*

Theorem 43 (**ANOVA*).** *Consider the following array of π-systems of $\mathcal{F}$-sets*

$$\begin{array}{ccc} \mathscr{A}_{1,1} & \mathscr{A}_{1,2} & \cdots \\ \mathscr{A}_{2,1} & \mathscr{A}_{2,2} & \cdots \\ \mathscr{A}_{3,1} & \mathscr{A}_{3,2} & \cdots \\ \vdots & \vdots & \ddots \end{array}$$

Each row may have a different number of columns (finite or infinite) and the number of rows may be finite or infinite. Let $\mathscr{R}_1, \mathscr{R}_2, \ldots$ denote the σ-fields generated by the rows: $\mathscr{R}_i := \sigma\langle\mathscr{A}_{i,1}, \mathscr{A}_{i,2}, \cdots\rangle$. Then the full collection $\{\mathscr{A}_{i,k}\}$ of π-systems are independent if and only if the following two statements hold:

1. *The π-systems within each row are independent;*
2. *The σ-fields generated by the rows, $\mathscr{R}_1, \mathscr{R}_2, \ldots$, are independent.*

Theorem 44 (**Independent binary digits).** *Let $H_n := \{w \in (0,1] : d_n(w) = 1\}$ where d_n denote the n^{th} binary (non-terminating) digit from the basic spinner model in Section 1. Then $H_1, H_2, \ldots$ are independent events under the model $P : \mathcal{B}_0^{(0,1]} \to [0,1]$ defined in Section 1.*

4.1 Borel-Cantelli, Fatou and 0-1 laws

Definition 29 (**Tail events).** *Let $\mathscr{A}_1, \mathscr{A}_2, \ldots$ be classes of $\mathcal{F}$-sets. Then*

$$\mathcal{T} := \bigcap_{m=1}^\infty \sigma\langle\mathscr{A}_n : n \geq m\rangle$$

is called the tail σ-field associated with $\mathscr{A}_1, \mathscr{A}_2, \ldots$. Moreover, any event $T \in \mathcal{T}$ is called a tail event.

Theorem 45 (**Kolmogorov's 0-1 Law).** *Suppose $\{\mathscr{A}_k\}_{k\in\mathcal{K}}$ are independent classes of $\mathcal{F}$-sets such that each $\mathscr{A}_k$ is also a π-system. Then for any tail event $T \in \mathcal{T}$ either $P(T) = 0$ or $P(T) = 1$.*

Definition 30 (*i.o.* **and** a.a.). *Let $A_1, A_2, \ldots$ be subsets of Ω. Then*

$$\{A_n \ \text{i.o.}_n\} := \bigcap_{m=1}^\infty \bigcup_{n=m}^\infty A_n =: \limsup_{n\to\infty} A_n$$
$$\{A_n \ \text{a.a.}_n\} := \bigcup_{m=1}^\infty \bigcap_{n=m}^\infty A_n =: \liminf_{n\to\infty} A_n$$

Theorem 46 (**1$^{\text{st}}$ Borel-Cantelli lemma).** *Let $A_1, A_2, \ldots$ be $\mathcal{F}$-sets. Then*

$$\sum_{n=1}^\infty P(A_n) < \infty \Longrightarrow P(A_n \ \text{i.o.}_n) = 0.$$

Theorem 47 (**Fatou for sets).** *Let $A_1, A_2, \ldots$ be $\mathcal{F}$-sets. Then*

$$\begin{aligned} P(A_n \ \text{a.a.}_n) &\leq \liminf_n P(A_n) \\ &\leq \limsup_n P(A_n) \leq P(A_n \ \text{i.o.}_n). \end{aligned}$$

Theorem 48 (**2$^{\text{nd}}$ Borel-Cantelli lemma).** *Let $A_1, A_2, \ldots$ be independent $\mathcal{F}$-sets. Then*

$$\sum_{n=1}^\infty P(A_n) = \infty \Longrightarrow P(A_n \ \text{i.o.}_n) = 1.$$

Exercise 21. *Let $\mathscr{A}_1, \ldots, \mathscr{A}_n$ be π-systems of $\mathcal{F}$-sets such that*

$$P\Big(\bigcap_{k=1}^n A_k\Big) = \prod_{k=1}^n P(A_k) \tag{28}$$

for each choice of $A_k \in \mathscr{A}_k$ for $k = 1, \ldots, n$. (a) Show by simple example that the $\mathscr{A}_k$'s need not be independent. (b) Show that the $\mathscr{A}_k$'s will be independent if for each k, Ω is the countable union of $\mathscr{A}_k$-sets. Hint: for fixed $A_2, \ldots, A_n$ use the following general inclusion-exclusion formula to show that (28) with A_1 replaced by any finite union of $\mathscr{A}_1$-sets. Here is the general inclusion-exclusion formula:

$$P\Big(\bigcup_{i=1}^n B_i\Big) = \sum_{m=1}^n (-1)^{m-1} \sum_{i_1 < i_2 < \cdots < i_m} P(B_{i_1} \cap B_{i_2} \cap \cdots \cap B_{i_m})$$

for any $\mathcal{F}$-sets $B_1, B_2, \ldots, B_n$.

Exercise 22. *(a) For each $k = 1, \ldots, n$ let $\mathcal{P}_k$ be a partition of Ω into countably many $\mathcal{F}$-sets. Show that the σ-fields $\sigma\langle\mathcal{P}_1\rangle, \ldots, \sigma\langle\mathcal{P}_n\rangle$ are independent if and only if (28) holds for each choice of A_k from $\mathcal{P}_k$ for $k = 1, \ldots, n$. (b) Use part (a) to show that $\mathcal{F}$-sets $A_1, \ldots, A_n$ are independent if and only if*

$$P\Big(\bigcap_{k=1}^n B_k\Big) = \prod_{k=1}^n P(B_k)$$

for each choice of B_k as A_k or A_k^c for $k = 1, \ldots, n$. (c) Use part (b) to show that the events $H_1, \ldots, H_n$ in Theorem 44 are independent.

Exercise 23. *Let $A_1, A_2, \ldots$ be $\mathcal{F}$-sets. Show that $P(A_n \text{ i.o.}_n) = 1$ if and only if $\sum_{n=1}^\infty P(A_n|A)$ diverges for every $\mathcal{F}$-set A of nonzero probability. Hint: show $P(A_n \text{ i.o.}_n) < 1 \iff \sum_{n=1}^\infty P(A_n|A) < \infty$ for some $\mathcal{F}$-set A with $P(A) > 0$.*

Exercise 24. *Let P and Q be probability measures on a σ-field $\mathcal{F}$ of subsets of a sample space Ω.*

- *P and Q are said to be* **singular**, *denoted $P \perp Q$, if and only if there exists a set $F \in \mathcal{F}$ such that*

$$P(F^c) = 0 = Q(F).$$

- *P is said to be* **absolutely continuous with respect to** *Q, denoted $P \ll Q$, if and only if*

$$P(F) = 0 \text{ for every } \mathcal{F}\text{-set } F \text{ for which } Q(F) = 0.$$

Show that

$$P \perp Q \iff \begin{bmatrix} \text{there exists } \mathcal{F}\text{-sets } F_1, F_2, \ldots \text{ such that} \\ P(F_n^c) \to 0 \text{ and } Q(F_n) \to 0 \text{ as } n \to \infty \end{bmatrix}$$

and

$$P \ll Q \iff \lim_{\delta \downarrow 0}\Big(\sup\{P(F)\colon F \in \mathcal{F} \text{ with } Q(F) \le \delta\}\Big) = 0.$$

4.2 Application: law of the iterated logarithm for coin flips

Section Assumption. *Throughout this section let $P : \mathcal{B}^{(0,1]} \to [0,1]$ be the probability model developed in Section 1 (and extended from the Carathéodory) for a uniform random number in $w \in (0,1]$. Also let s_n be defined as in Section 1.*

To motivate the law of the iterated logarithm lets start by discussing the difference between the weak law and strong law of large numbers.

$$\textbf{Weak law}: \quad \lim_{n\to\infty} P\Big(\big|\frac{s_n}{n}\big| < \epsilon\Big) = 1, \text{ for all } \epsilon > 0;$$

$$\textbf{Strong law}: \quad P\Big(\lim_{n\to\infty} \frac{s_n}{n} = 0\Big) = 1.$$

In some sense the difference can be explained as follows. The weak law fixes each n then analyzes the ensemble of $s_n(\omega)/n$ over $\omega \in (0,1]$. In particular, the weak law says that for large n it becomes increasingly rare to find ω's which satisfy $|s_n(\omega)/n| \ge \epsilon$. Conversely the strong law fixes each $\omega \in (0,1]$ and analyzes the ensemble of $s_n(\omega)/n$ over n. In particular for almost all ω, $s_n(\omega)/n \to 0$ as $n \to \infty$.

Now lets make a similar analogy with the central limit theorem and the law of the iterated logarithm. From the strong law we know that $s_n(\omega)/n$ converges to 0 for nearly every ω. We can then ask: at what rate? In particular, can we find a smaller denominator than n, call it ℓ_n, so that s_n/ℓ_n doesn't converge to zero. An answer is given by the central limit theorem

$$\textbf{CLT}: \quad \lim_{n\to\infty} P\Big(\frac{s_n}{\sqrt{n}} < x\Big) = \Phi(x)$$

where $\Phi(x) = P(Z \le x)$ and Z is a standard normal random variable. Notice two things. First, this suggests that the $s_n(\omega)/\sqrt{n}$ reaches up to ∞ and down to $-\infty$ for different values of ω and n. In particular for every cut-off M, there exists n large enough so that $P\Big(s_n/\sqrt{n} \ge M\Big) \approx 1 - \Phi(M) > 0$ to arbitrary precision. Also, the central limit theorem is similar to the weak law in that it fixes each n then analyzes the ensemble of $s_n(\omega)/\sqrt{n}$ over $\omega \in (0,1]$. The question then becomes, can one find an analogous form of the strong law such that for each fixed ω one analyzes the ensemble rate of $s_n(\omega)$ as $n \to \infty$. The law of the iterated logarithm gives the right rate

$$\textbf{LIL}: \quad P\Big(\limsup_{n\to\infty} \frac{s_n}{\sqrt{2n\log\log n}} = 1\Big) = 1.$$

Another way to think about the rate $\sqrt{2n\log\log n}$ is the effect due to the correlation of between s_n across different values of n. The expected maximum of n independent standard Gaussian random variables behaves as $\sqrt{2\log n}$. That maximum occurs uniformly on $\{1, 2, \ldots, n\}$ and is therefore is expected to occur at index $n/2$. If $s_n/\sqrt{n}$ was not correlated across n, the central limit theorem might suggest that the maximum of $s_k/\sqrt{k}$ over $k \in \{1, 2 \ldots, n\}$ behaves on the order of $\sqrt{2\log n}$. This loosely suggests the maximum of s_k over $k \in \{1, 2 \ldots, n\}$ behaves on the order $\sqrt{n\log n}$. Now, in some sense, the LIIL says that the correlation across n will dampen the maximum excursions to be at most $\sqrt{n\log\log n}$.

Lemma 1 (Half of large deviation result). *For all $n\in\mathbb{N}$ and $x>0$ one has*

$$P\big(s_n/\sqrt{n}\geq x\big)\leq \exp\Big(-\frac{x^2}{2}\Big)$$

Proof. This was established in exercise 2. □

The other half of the large deviation result we need is Lemma 2, below. Combined these two lemmas give us good approximations to $P\big(s_n/\sqrt{n}\geq x_n\big)$ for large-ish values of x_n: large compared to 0 but still small compared to $\sqrt{n}$ (if x_n was larger then $\sqrt{n}$ then $P(s_n/\sqrt{n}>x_n)=0$). This is the key for deriving the Law of the Iterated Logarithm. Also note that Lemma 1 was proved as an exercise but it is typically established using Markov's inequality, the moment generating function and the fact that $\frac{e^x+e^{-x}}{2}\leq\exp\big(x^2/2\big)$.

Lemma 2 (Other half of large deviation result). *If the sequence $\{x_n\}_{n\in\mathbb{N}}$ satisfies $0\leq x_n\to\infty$ and $x_n/\sqrt{n}\to 0$ as $n\to\infty$ then*

$$P\big(s_n/\sqrt{n}\geq x_n\big)\geq \exp\Big(-\frac{x_n^2}{2}(1+o(1))\Big).$$

Proof. The general idea is to use the fact that $s_n = 2\big(\sum_{k=1}^n d_k\big)-n$ where $\sum_{k=1}^n d_k\sim \mathrm{Bin}(n,1/2)$. Therefore we can write $P(s_n\geq\sqrt{n}x_n)$ as a sum $\sum_{i\in\mathcal{I}_n}P(s_n=i)$ where i is the set of integers greater than $\sqrt{n}x_n$ and less than or equal to n. In fact, since we are trying to construct a lower bound we are free to discard terms in $\mathcal{I}_n$ which will give

$$P(s_n\geq\sqrt{n}x_n)\geq\sum_{i\in\mathcal{I}_n}P(s_n=i).$$

The main problem is how to find $\mathcal{I}_n$ so that the right hand side is $\exp\big(-\frac{x_n^2}{2}(1+o(1))\big)$.

Lets start by getting some idea of how many integers we should include in $\mathcal{I}_n$ by analysing how $P(s_n=i)$ behaves when $i\approx\sqrt{n}x_n$. To make things a bit more precise let i_n be a sequence of integers depending on n such that $i_n\to\infty$ but $i_n/n\to 0$.

$$\begin{aligned}
P(s_n=i_n) &= P\Big(\sum_{k=1}^n d_k=(i_n+n)/2\Big)\\
&=\binom{n}{(i_n+n)/2}\frac{1}{2^n},\quad \text{if } (i_n+n)/2 \text{ is an integer}\\
&=\frac{n!}{\frac{i_n+n}{2}!\,\frac{n-i_n}{2}!}\frac{1}{2^n}\\
&=\frac{2}{\sqrt{2\pi n}}[1+o(1)]\exp\Big(-\frac{(1+o(1))i_n^2}{2n}\Big) \qquad (29)
\end{aligned}$$

To see why (29) is true notice that by Stirling's formula we have $n!=(1+o(1))\sqrt{2\pi n}n^ne^{-n}$. Therefore

$$\begin{aligned}
\frac{n!}{\frac{i_n+n}{2}!\,\frac{n-i_n}{2}!}\frac{1}{2^n} &= \frac{(1+o(1))}{(1+o(1))^2}\\
&\quad\times\frac{\sqrt{2\pi n}}{\sqrt{2\pi(n+i_n)/2}\sqrt{2\pi(n-i_n)/2}}\\
&\quad\times\frac{n^n}{(n+i_n)^{(n+i_n)/2}(n-i_n)^{(n-i_n)/2}}\\
&\quad\times\frac{e^{-n}}{e^{-(n+i_n)/2}e^{-(n-i_n)/2}}\\
&=\frac{(1+o(1))}{(1+o(1))^2}\\
&\quad\times\underbrace{\frac{1}{\sqrt{2\pi(n+i_n)(n-i_n)/(4n)}}}_{=:I}\\
&\quad\times\underbrace{\Big(\frac{n}{n+i_n}\Big)^{(n+i_n)/2}\Big(\frac{n}{n-i_n}\Big)^{(n-i_n)/2}}_{=:II}
\end{aligned}$$

Notice

$$\begin{aligned}
I &= \frac{1}{\sqrt{2\pi(n+i_n)(n-i_n)/(4n)}}\\
&=\frac{2}{\sqrt{2\pi n}}\times\frac{1}{\sqrt{(n+i_n)(n-i_n)/(n^2)}}\\
&=\frac{2}{\sqrt{2\pi n}}\times\frac{1}{\sqrt{1-(i_n/n)^2}}=\frac{2}{\sqrt{2\pi n}}(1+o(1)).
\end{aligned}$$

Secondly notice that $(1+x)\log(1+x)=x+\frac{1}{2}x^2+O(x^3)$ as $x\to 0$. Therefore

$$\begin{aligned}
\log II &= -\frac{1}{2}\big[(n+i_n)\log\big(1+\frac{i_n}{n}\big)+(n-i_n)\log\big(1-\frac{i_n}{n}\big)\big]\\
&=-\frac{n}{2}\Big[\frac{i_n}{n}+\frac{1}{2}\frac{i_n^2}{n^2}-\frac{i_n}{n}+\frac{1}{2}\frac{i_n^2}{n^2}+O(i_n^3/n^3)\Big]\\
&=-\frac{n}{2}\Big[\frac{i_n^2}{n^2}+O(i_n^3/n^3)\Big]=-\frac{i_n^2}{2n}\big[1+\underbrace{O(i_n/n)}_{o(1)}\big].
\end{aligned}$$

To finish notice that $\frac{(1+o(1))^2}{(1+o(1))^2}=[1+o(1)]$ which implies (29).

Now, looking at (29) it is clear that we want $\mathcal{I}_n$ to contain about $\sqrt{2\pi n}$ terms that are near $\sqrt{n}x_n$ (so that we can apply (29)). In particular $\mathcal{I}_n$ denote the set of indices between $\sqrt{n}x_n$ and $\sqrt{n}x_n+2\sqrt{\pi n}$ which have the same parity at n (i.e. that $(i_n+n)/2$ is an integer). Also let i_n be the maximum integer in $\mathcal{I}_n$, which implies $i_n=\sqrt{n}x_n+2\sqrt{\pi n}+O(1)$. Now

$$\begin{aligned}
&P(s_n\geq\sqrt{n}x_n)\\
&\quad\geq\sum_{i\in\mathcal{I}_n}P(s_n=i)\\
&\quad\geq[\#\mathcal{I}_n]P(s_n=i_n),\ \text{since}\ \binom{n}{(i+n)/2}\geq\binom{n}{(i_n+n)/2}
\end{aligned}$$

$$= [\sqrt{\pi n} + O(1)]P(s_n = i_n)$$
$$\geq \sqrt{2}\exp\left(-\frac{(1+o(1))i_n^2}{2n}\right), \quad \text{by (29)}$$
$$\geq \exp\left(-\frac{(1+o(1))i_n^2}{2n}\right)$$
$$= \exp\left(-\frac{(1+o(1))\big[\sqrt{n}x_n + 2\sqrt{\pi n} + O(1)\big]^2}{2n}\right)$$
$$= \exp\left(-\frac{x_n^2}{2}(1+o(1))\right), \quad \text{since } x_n \to \infty.$$

□

Lemma 3 (Maximal inequality). *For all $n \in \mathbb{N}$ and every nonnegative integer c*

$$P\big(\max_{1\leq j\leq n} s_j \geq c\big) \leq 2P(s_n \geq c).$$

Proof. First write

$$P\big(\max_{1\leq j\leq n} s_j \geq c\big) = P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n \geq c\big) + P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n < c\big)$$
$$= P\big(s_n \geq c\big) + P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n < c\big).$$

Therefore all that remains is to show $P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n < c\big) \leq P\big(s_n \geq c\big)$. Start by segmenting the event $\{\max_{1\leq j\leq n} s_j \geq c\}$ corresponding to the first indicie j for which $s_j = c$ (this must occur when $\max_{1\leq j\leq n} s_j \geq c$ since s_n goes up or down with jumps of size 1 and c is a nonnegative integer). In particular, write

$$\{\max_{1\leq j\leq n} s_j \geq c\} = \bigcup_{j=1}^{n} \underbrace{\{s_1 < c,\, \dots,\, s_{j-1} < c,\, s_j = c\}}_{=:F_j}$$

Now

$$\{\max_{1\leq j\leq n} s_j \geq c\} \cap \{s_n < c\}$$
$$= \bigcup_{j=1}^{n} F_j \cap \{s_n < c\}$$
$$= \bigcup_{j=1}^{n} \underbrace{F_j \cap \{s_n - s_j < 0\}}_{\text{disjoint since the } F_j\text{'s are}}, \quad \text{since } \omega \in F_j \text{ implies } s_j(\omega) = c.$$

Now notice two things. First $P(s_n - s_j < 0) = P(s_n - s_j > 0)$ by symmetry. Secondly, since $\{s_n - s_j < 0\} \in \sigma\langle z_{j+1}, \dots, z_n\rangle$ and $F_j \in \sigma\langle z_1, \dots, z_j\rangle$, the event $\{s_n - s_j < 0\}$ is independent of F_j (by ANOVA). Therefore

$$P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n < c\big)$$
$$= \sum_{j=1}^{n} P\big(F_j \cap \{s_n - s_j < 0\}\big)$$
$$= \sum_{j=1}^{n} P\big(F_j\big)P\big(s_n - s_j < 0\big) \text{ by independence}$$
$$= \sum_{j=1}^{n} P\big(F_j\big)P\big(s_n - s_j > 0\big) \text{ by symmetry}$$
$$= \sum_{j=1}^{n} P\big(F_j \cap \{s_n - s_j > 0\}\big) \text{ by independence again}$$
$$= \sum_{j=1}^{n} P\big(F_j \cap \{s_n > c\}\big) \text{ since } \omega \in F_j \text{ implies } s_j(\omega) = c$$
$$= P\big(\max_{1\leq j\leq n} s_j \geq c,\, s_n > c\big)$$
$$\leq P\big(s_n > c\big)$$
$$\leq P\big(s_n \geq c\big).$$

Therefore $P\big(\max_{1\leq j\leq n} s_j \geq c\big) \leq 2P\big(s_n \geq c\big)$. □

Theorem 49 (Law of the iterated logarithm for coin flips).

1. $P\Big[\limsup_{n\to\infty} \dfrac{s_n}{\sqrt{2n\log\log n}} = 1\Big] = 1$*;*
2. $P\Big[\liminf_{n\to\infty} \dfrac{s_n}{\sqrt{2n\log\log n}} = -1\Big] = 1.$

Proof. To make the formulas more readable let $\ell_n := \sqrt{2n\log\log n}$. Notice first that

$$\{\limsup_n \frac{s_n}{\ell_n} = 1\} = \bigcap_{\epsilon\in(0,1)\cap\mathbb{Q}} \{s_n/\ell_n > (1-\epsilon) \text{ i.o.}_n\} \cap \{s_n/\ell_n < (1+\epsilon) \text{ a.a.}_n\}.$$

This implies that $\{\limsup_n (s_n/\ell_n) = 1\}$ and $\{\liminf_n (s_n/\ell_n) = -1\}$ are Borel measurable. Secondly notice that by symmetry we have

$$P\big[\limsup_{n\to\infty}(s_n/\ell_n) = 1\big] = P\big[\liminf_{n\to\infty}(-s_n/\ell_n) = -1\big] = P\big[\liminf_{n\to\infty}(s_n/\ell_n) = -1\big].$$

We can also simply our proof by using countable sub-additivity

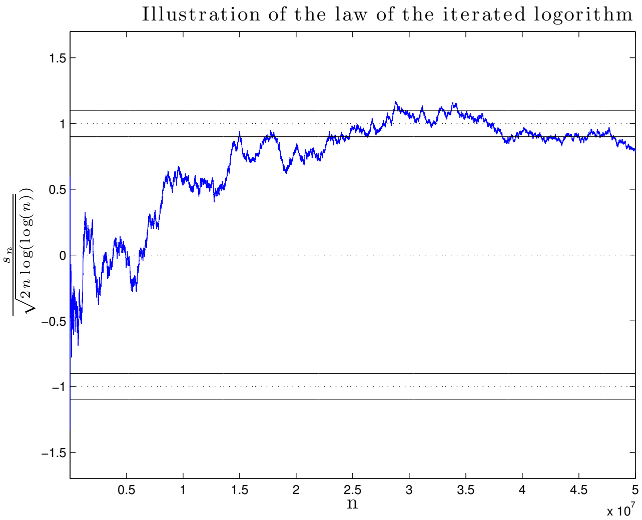
$$P\big[\big\{\limsup_{n\to\infty}(s_n/\ell_n) = 1\big\}^c\big] \leq \sum_{\epsilon\in(0,1)\cap\mathbb{Q}} P[\{s_n/\ell_n \leq (1-\epsilon) \text{ a.a.}_n\}] + \sum_{\epsilon\in(0,1)\cap\mathbb{Q}} P[\{s_n/\ell_n \geq (1+\epsilon) \text{ i.o.}_n\}].$$

Therefore the proof will follow by establishing Lemmas 4 and 5 which state that for all $\epsilon > 0$

$$P[s_n/\ell_n \geq (1+\epsilon) \text{ i.o.}_n] = 0$$
$$P[s_n/\ell_n > (1-\epsilon) \text{ i.o.}_n] = 1$$

□

Theorem 49 is interesting for a number of reasons. First, it gives a very detailed analysis of the Central Limit Theorem. Second, it shows the power of the Borel-Cantelli lemmas. Third, it is one of those theorems in probability which is extremely hard to see in simulations: $\sqrt{2n \log\log n}$ grows to slow for modern computers to probe. The following simulation is an attempt at illustrating Theorem 186 but we still don't see the required fluctuation from $+1$ to -1.



Lemma 4. *For all* $\epsilon > 0$

$$P[s_n/\ell_n \geq (1+\epsilon) \ i.o._n] = 0$$

Proof. The obvious strategy is to use the first Borel-Cantelli lemma. In particular, it would be nice if we could show

$$\sum_{n=1}^{\infty} P[s_n/\ell_n \geq (1+\epsilon)] < \infty$$

which would give us the lemma. By the first half of the large deviation result we know $P[s_n/\ell_n \geq (1+\epsilon)]$ converge to zero as $n \to \infty$. Unfortunately, they do not converge fast enough for the first Borel-Cantelli. Taking a sub-sequence will get something summable but we would need to show that the events are well behaved between the sub-sequence. Fortunatly we can do this since the sets $\{s_n/\ell_n \geq (1+\epsilon)\}$ overlap a lot. The strategy is to group the events $\{s_n/\ell_n \geq (1+\epsilon)\}$ by unioning them into blocks, then apply Borel-Cantelli on the blocks. This will be sufficient since the blocks occur infinity often if and only if the events $s_n/\ell_n \geq (1+\epsilon)$ occur infinity often. Controlling the probability of the blocks is done with the maximal inequality.

Let the k^{th} block be defined

$$B_k := \bigcup_{j=n_{k-1}}^{n_k} \{s_j/\ell_j \geq (1+\epsilon)\}$$

where n_k is a (yet to be determined) subsequence. We use maximal inequality to bound $P[B_k]$ as follows

$$\begin{aligned}
P[B_k] &\leq P[\max_{n_{k-1}\leq j\leq n_k} s_j \geq (1+\epsilon) \min_{n_{k-1}\leq j\leq n_k} \ell_j] \\
&\leq P[\max_{n_{k-1}\leq j\leq n_k} s_j \geq (1+\epsilon)\ell_{n_{k-1}}] \\
&\leq P[\max_{j\leq n_k} s_j \geq (1+\epsilon)\ell_{n_{k-1}}] \\
&\leq P[\max_{j\leq n_k} s_j \geq \lceil (1+\epsilon)\ell_{n_{k-1}} \rceil], \quad \text{since } s_j \in \mathbb{Z} \\
&\leq 2P[s_{n_k} \geq \lceil (1+\epsilon)\ell_{n_{k-1}} \rceil], \quad \text{maximal ineq} \\
&\leq 2P[s_{n_k} \geq (1+\epsilon)\ell_{n_{k-1}}] \\
&\leq \exp\big(-\tfrac{1}{2}(1+\epsilon)^2 \ell_{n_{k-1}}^2/n_k\big), \quad \text{half of large deviation} \\
&\leq \exp\big(-(1+\epsilon)^2 n_{k-1} \log\log n_{k-1}/n_k\big) \\
&= \left(\frac{1}{\log n_{k-1}}\right)^{(1+\epsilon)^2 \frac{n_{k-1}}{n_k}}.
\end{aligned}$$

Now we just find n_k which makes the last term summable over k. If $n_k \approx \theta^k$ one gets

$$\left(\frac{1}{\log n_{k-1}}\right)^{(1+\epsilon)^2 \frac{n_{k-1}}{n_k}} \approx \left(\frac{1}{(k-1)\log\theta}\right)^{(1+\epsilon)^2 \frac{1}{\theta}}$$

which is summable if $(1+\epsilon)^2 \frac{1}{\theta} > 1$. We also need that $\theta > 1$ since we need $n_k \to \infty$ as $k \to \infty$. Luckily there does exist such a θ for which $(1+\epsilon)^2 > \theta > 1$. $\square$

Lemma 5. *For all* $\epsilon > 0$

$$P[s_n/\ell_n > (1-\epsilon) \ i.o._n] = 1$$

Proof. In the previous lemma we presented a technique to adjust the first Borel-Cantelli lemma in the case the summability condition doesn't hold. For this lemma we want to use the second Borel-Cantelli lemma but, again, it doesn't directly apply since the condition that the events $s_n/\ell_n > (1-\epsilon)$ are independent does not hold. Here is a generic technique to get around this obstacle. Find subsequence n_k and subsets

$$I_k \subset \{s_{n_k}/\ell_{n_k} > (1-\epsilon)\}$$

such that I_k's are independent and $\sum_k P[I_k] = \infty$ so that $P[I_k \ i.o._k] = 1$ (which would then give the lemma). Unfortunately, even this doesn't work. What ends up working is to find two sets A_k I_k such that

$$A_k \cap I_k \subset \{s_{n_k}/\ell_{n_k} > (1-\epsilon)\} \tag{30}$$

$$I_k \text{ are independent and } \textstyle\sum_k P[I_k] = \infty \tag{31}$$

$$A_k \text{ are not independent but } P[A_k \ a.a._k] = 1. \tag{32}$$

To see why this is sufficient notice that (31) implies $P[I_k \ i.o._k] = 1$ by the second Borel-Cantelli lemma. Then

$$\begin{aligned}
&P[A_k \ a.a._k] = 1 \text{ and } P[I_k \ i.o._k] = 1 \\
&\Longrightarrow P[A_k \cap I_k \ i.o._k] = 1 \\
&\overset{(30)}{\Longrightarrow} P[s_{n_k}/\ell_{n_k} > (1-\epsilon) \ i.o._k] = 1
\end{aligned}$$

Therefore (30), (31) and (32) are sufficient to establish the lemma.

Figuring out how to define I_k and A_k are the tricky parts. The intuition is that if your going to get independent events you need to look at increments of s_n. Define

$$I_k := \{s_{n_k} - s_{n_{k-1}} \geq (1-\epsilon/2)\ell_{n_k}\}$$
$$A_k := \{s_{n_{k-1}} > -(\epsilon/2)\ell_{n_k}\}.$$

Clearly $I_k \cap A_k \subset \{s_{n_k}/\ell_{n_k} > (1-\epsilon)\}$ so that (30) holds. Moreover, the I_k's are independent. To show (32)

$$\begin{aligned}
P[A_k \text{ a.a.}_k] &= P\Big[s_{n_{k-1}} > -(\epsilon/2)\ell_{n_k} \text{ a.a.}_k\Big] \\
&= P\Big[s_{n_{k-1}} < (\epsilon/2)\ell_{n_k} \text{ a.a.}_k\Big], \quad \text{by symmetry} \\
&= P\Big[\frac{s_{n_{k-1}}}{\ell_{n_{k-1}}} < (\epsilon/2)\frac{\ell_{n_k}}{\ell_{n_{k-1}}} \text{ a.a.}_k\Big] \\
&= 1 - P\Big[\frac{s_{n_{k-1}}}{\ell_{n_{k-1}}} \geq (\epsilon/2)\frac{\ell_{n_k}}{\ell_{n_{k-1}}} \text{ i.o.}_k\Big] \\
&= 1, \text{ by Lemma 4 if } \frac{\ell_{n_k}}{\ell_{n_{k-1}}} \to \infty.
\end{aligned}$$

To show (31) notice

$$\begin{aligned}
P[I_k] &= P\Big[s_{n_k} - s_{n_{k-1}} \geq (1-\epsilon/2)\ell_{n_k}\Big] \\
&= P\Big[s_{n_k - n_{k-1}} \geq (1-\epsilon/2)\ell_{n_k}\Big] \\
&\geq P\Big[\frac{s_{n_k - n_{k-1}}}{\sqrt{n_k - n_{k-1}}} \geq \frac{(1-\epsilon/2)\ell_{n_k}}{\sqrt{n_k - n_{k-1}}}\Big] \\
&\geq \exp\Big(-\frac{1}{2}\Big[\frac{(1-\epsilon/2)\ell_{n_k}}{\sqrt{n_k - n_{k-1}}}\Big]^2 (1+o(1))\Big), \qquad (33)
\end{aligned}$$

by Lemma 2 if:

$\frac{\ell_{n_k}}{\sqrt{n_k - n_{k-1}}} \to \infty$ and

$\frac{1}{\sqrt{n_k - n_{k-1}}}\frac{\ell_{n_k}}{\sqrt{n_k - n_{k-1}}} \to 0$ and

$n_k - n_{k-1} \to \infty.$

To finish the proof of (32) and (31) we need to find $n_k \to \infty$ such that ℓ_{n_k} satisfies the above conditions and the sum of (33) diverges.

A subsequence of the form $n_k := \lfloor \exp(k^\theta) \rfloor$ will work. To check the conditions notice

$$\begin{aligned}
\frac{\ell_{n_k}}{\ell_{n_{k-1}}} &\sim \frac{\sqrt{2\exp(k^\theta)\log k^\theta}}{\sqrt{2\exp((k-1)^\theta)\log(k-1)^\theta}} \\
&= \exp\Big(\frac{k^\theta - (k-1)^\theta}{2}\Big)\frac{\log k}{\log(k-1)} \\
&= \exp\Big(\frac{\theta(k^*)^{\theta-1}}{2}\Big)(1+o(1)), \quad \text{where } k-1 \leq k^* \leq k \\
&\longrightarrow \infty, \quad \text{if } \theta > 1.
\end{aligned}$$

Also

$$\begin{aligned}
n_k - n_{k-1} &\sim \exp(k^\theta) - \exp((k-1)^\theta) \\
&= \theta(k^*)^{(\theta-1)}\exp((k^*)^\theta), \quad \text{where } k-1 \leq k^* \leq k \\
&\longrightarrow \infty, \quad \text{if } \theta > 0.
\end{aligned}$$

And therefore

$$\begin{aligned}
\frac{\ell_{n_k}}{\sqrt{n_k - n_{k-1}}} &\sim \frac{\sqrt{2\exp(k^\theta)\log k^\theta}}{\sqrt{\exp(k^\theta) - \exp((k-1)^\theta)}} \\
&= \frac{\sqrt{2\theta \log k}}{\sqrt{1 - \exp((k-1)^\theta - k^\theta)}} \\
&= \frac{\sqrt{2\theta\log k}}{\sqrt{1-o(1)}} \\
&\longrightarrow \infty.
\end{aligned}$$

Clearly we now have that $\frac{\ell_{n_k}}{n_k - n_{k-1}} \longrightarrow 0$. Finally we need to show that the sum of (33) over k diverges. The individual terms are

$$\begin{aligned}
&\exp\Big(-\frac{1}{2}\Big[\frac{(1-\epsilon/2)\ell_{n_k}}{\sqrt{n_k - n_{k-1}}}\Big]^2(1+o(1))\Big) \\
&\sim \exp\Big(-\frac{(1-\epsilon/2)^2}{2}2\theta\log k\Big) \\
&= \exp\Big(-(1-\epsilon/2)^2\theta\log k\Big) \\
&= k^{-(1-\epsilon/2)^2\theta}
\end{aligned}$$

the sum of the above terms over k diverges if $(1-\epsilon/2)^\theta < 1$, i.e. if $\theta < \frac{1}{(1-\epsilon)^2}$. Now putting all the conditions on θ together says that we simply need to choose θ such that

$$1 < \theta < \frac{1}{(1-\epsilon/2)^2}.$$

□

Part II
Integration

5 Measurable functions

Definition 31 (Measurable functions). *If $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are measurable spaces then $f\colon \Omega_1 \to \Omega_2$ is said to be measurable between $\mathcal{F}_1$ and $\mathcal{F}_2$ (written $f ⓜ \mathcal{F}_1/\mathcal{F}_2$) if and only if*

$$f^{-1}(A) \in \mathcal{F}_1, \quad \forall A \in \mathcal{F}_2$$

where $f^{-1}(A) := \{w \in \Omega_1 : f(w) \in A\}$.

Theorem 50 (Basic facts about pull backs). *Let $f\colon \Omega_1 \to \Omega_2$. Let $A, A_1, A_2, \ldots \subset \Omega_2$. Then*

- $f^{-1}(\Omega_2) = \Omega_1$
- $f^{-1}(\varnothing) = \varnothing$
- $f^{-1}(\Omega_2 - A) = \Omega_1 - f^{-1}(A)$
- $f^{-1}(\cup_i A_i) = \cup_i f^{-1}(A_i)$
- $f^{-1}(\cap_i A_i) = \cap_i f^{-1}(A_i)$.

Proof. These are very easy to check. For example to see why $f^{-1}(\cup_i A_i) \subset \cup_i f^{-1}(A_i)$ notice that

$$\begin{aligned} \omega \in f^{-1}(\cup_i A_i) &\Longrightarrow f(\omega) \in \cup_i A_i \\ &\Longrightarrow f(\omega) \in A_i \text{ for some } i \\ &\Longrightarrow \omega \in f^{-1}(A_i) \text{ for some } i \\ &\Longrightarrow \omega \in \cup_i f^{-1}(A_i). \end{aligned}$$

The other arguments are exactly similar. □

Theorem 51 (Generators are enough). *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces where $\mathcal{F}_2$ is generated by some class $\mathcal{A} \subset 2^{\Omega_2}$ (i.e. $\mathcal{F}_2 = \sigma\langle\mathcal{A}\rangle$). If $f : \Omega_1 \to \Omega_2$ then*

$$f ⓜ \mathcal{F}_1/\mathcal{F}_2 \Longleftrightarrow f^{-1}(A) \in \mathcal{F}_1, \ \forall A \in \mathcal{A}.$$

Corollary 3 (Monotone real maps are measurable). *Any monotone map $f\colon \mathbb{R} \to \mathbb{R}$ is measurable $\mathcal{B}^{\mathbb{R}}/\mathcal{B}^{\mathbb{R}}$.*

Corollary 4 (Continuous real maps are measurable). *Any continuous map $f\colon \mathbb{R}^d \to \mathbb{R}^k$ is measurable $\mathcal{B}^{\mathbb{R}^d}/\mathcal{B}^{\mathbb{R}^k}$.*

Theorem 52 (Composition of ⓜ functions is ⓜ). *Let $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$ and $(\Omega_3, \mathcal{F}_3)$ be measurable spaces. Suppose f and g are functions sending $\Omega_1 \xrightarrow{f} \Omega_2 \xrightarrow{g} \Omega_3$. If $f ⓜ \mathcal{F}_1/\mathcal{F}_2$ and $g ⓜ \mathcal{F}_2/\mathcal{F}_3$ then $g \circ f ⓜ \mathcal{F}_1/\mathcal{F}_3$.*

Corollary 5 (Just check that each coordinate is ⓜ). *Let $(\Omega, \mathcal{F})$ be a measurable space and $f : \Omega \to \mathbb{R}^d$. Let $f = (f_1, \ldots, f_d)$ decompose f into the coordinate functions (so that $f_k : \Omega \to \mathbb{R}$). Then*

$$f ⓜ \mathcal{F}/\mathcal{B}^{\mathbb{R}^d} \Longleftrightarrow f_k ⓜ \mathcal{F}/\mathcal{B}^{\mathbb{R}}, \text{ for each } k.$$

Theorem 53 (Scissors and paste). *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces and $f\colon \Omega_1 \to \Omega_2$. In addition, suppose there exists $\mathcal{F}_1$-sets $A_1, A_2, \ldots$ such that $\Omega_1 = \cup_{k=1}^\infty A_k$. Then*

$$f ⓜ \mathcal{F}_1/\mathcal{F}_2 \Longleftrightarrow f\big|_{A_k} ⓜ (\mathcal{F}_1 \cap A_k)/\mathcal{F}_2, \text{ for each } k.$$

Corollary 6 (Piecewise-continuous real maps are measurable). *Any map $f\colon \mathbb{R}^d \to \mathbb{R}^k$ which is piecewise continuous on each piece of a countable-measurable-partition of $\mathbb{R}^d$ is measurable $\mathcal{B}^{\mathbb{R}^d}/\mathcal{B}^{\mathbb{R}^k}$.*

Theorem 54 (Metric-continuous functions are measurable). *Suppose Ω_1 and Ω_2 are metric spaces and f is a function mapping Ω_1 into Ω_2. If there exists $\mathcal{B}^{\Omega_1}$ sets $A_1, A_2, \ldots$ such that $\Omega_1 = \cup_{k=1}^\infty A_k$ and $f\big|_{A_k}$ is continuous (with respect to the induced metrics) on each A_k then f is measurable $\mathcal{B}^{\Omega_1}/\mathcal{B}^{\Omega_2}$.*

Theorem 55 (Just check Borel ⓜ on the range). *Let $(\Omega_1, \mathcal{F}_1)$ be a measurable space and Ω_2 be a metric space. Suppose f is a function which maps Ω_1 into $\Omega_2^o \subset \Omega_2$. Then*

$$f ⓜ \mathcal{F}_1/\mathcal{B}^{\Omega_2^o} \Longleftrightarrow f ⓜ \mathcal{F}_1/\mathcal{B}^{\Omega_2}$$

where the metric used to define $\mathcal{B}^{\Omega_2^o}$ is the one induced by the metric on Ω_2.

Definition 32 (Nomenclature for the extended reals).

- *$\mathcal{B}$ is shorthand notation for $\mathcal{B}^{\bar{\mathbb{R}}}$.*
- *If $(\Omega, \mathcal{F})$ is a measurable space and $f : \Omega \to \bar{\mathbb{R}}$ we use the nomenclature 'ⓜ $\mathcal{F}$' or just 'measurable $\mathcal{F}$' as short hand for ⓜ $\mathcal{F}/\mathcal{B}$.*
- *We say that a function f is 'Borel measurable' or just 'measurable' if $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ and f is ⓜ $\mathcal{B}^{\mathbb{R}^d}/\mathcal{B}$.*
- *We say that a function f is 'Lebesque measurable' if $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ and f is ⓜ $\overline{\mathcal{B}^{\mathbb{R}^d}}/\mathcal{B}$.*

Definition 33 (Defining algebraic operations with ∞).

- *$\infty + c := \infty$ for any $-\infty < c \leq \infty$.*
- *$\infty \cdot 0 = 0 \cdot \infty := 0$.*
- *$\infty \cdot \infty := \infty$.*
- *$\frac{c}{\infty} := 0$ when $c \in \mathbb{R}$.*
- *$\frac{c}{0}$, $\frac{\pm\infty}{\pm\infty}$, $\infty - \infty$ and $-\infty + \infty$ are not defined.*

Theorem 56 (Closure theorem for ⓜ functions). *If $(\Omega, \mathcal{F})$ is a measurable space then*

1. *If f and g are ⓜ $\mathcal{F}$ functions then cf (c is a constant), $f+g$, fg, f/g, $f \vee g$, $f \wedge g$, f^+, f^- are each ⓜ $\mathcal{F}$, provided the composite function are defined at every $w \in \Omega$.*
2. *If $f_1, f_2, \ldots$ ⓜ $\mathcal{F}$ functions then $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, $\liminf_n f_n$ are each ⓜ $\mathcal{F}$.*

Definition 34 (**Simple functions).** *Let $(\Omega, \mathcal{F})$ denote a measurable space. Then any function $f\colon \Omega \to \bar{\mathbb{R}}$ which is ⓜ$\mathcal{F}$ and has a finite range is called a simple function.*

Definition 35 (**Characterization of simple functions).** *Let $(\Omega, \mathcal{F})$ denote a measurable space and suppose $f\colon \Omega \to \bar{\mathbb{R}}$. Then f is a simple function if and only if there exists a finite partition of Ω into disjoint $\mathcal{F}$-sets $A_1, \ldots, A_n$ and a finite list of extended real numbers $c_1, \ldots, c_n$ such that $f = \sum_{k=1}^n c_k I_{A_k}$.*

Definition 36 ($\mathscr{N}_s$ **and** $\mathscr{N}$**).** *Let $(\Omega, \mathcal{F})$ denote a measurable space. Then*

- *$\mathscr{N}_s$ denotes the set of non-negative simple functions $f : \Omega \to \bar{\mathbb{R}}$.*
- *$\mathscr{N}$ denotes the set of non-negative functions $f : \Omega \to \bar{\mathbb{R}}$ which are ⓜ$\mathcal{F}$.*

Theorem 57 (**The structure theorem).** *Let $(\Omega, \mathcal{F})$ be a measurable space and suppose $f\colon \Omega \to \bar{\mathbb{R}}$ is ⓜ$\mathcal{F}$. Then*

1. *There exists bounded simple functions $f_1, f_2, \ldots$ such that $\lim_n f_n(w) = f(w)$ for each $w \in \Omega$.*
2. *If, in addition, $f \in \mathscr{N}$ then there exists bounded $f_1, f_2, \ldots \in \mathscr{N}_s$ such that $f_n(w) \uparrow f(w)$ for each $w \in \Omega$.*

Exercise 25. *Show that Corollary 5 holds for functions mapping into $\bar{\mathbb{R}}^d$.*

Exercise 26. *Let $(\Omega, \mathcal{F})$ be a measure space and let $f_0, f_1, \ldots$ be an infinite sequence of $\mathcal{F}$-measurable functions of Ω. Show that the radius R of convergence of the random power series $\sum_{k=0}^\infty f_k x^k$ is an $\mathcal{F}$-measurable function of Ω.*

Exercise 27. *Give an example of two measurable spaces $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, a ⓜ$\mathcal{F}_1/\mathcal{F}_2$ mapping $f : \Omega_1 \to \Omega_2$, and an event $B \in \mathcal{F}_1$ such that $f(B) \notin \mathcal{F}_2$.*

5.1 Application: Random variables and distribution functions

Theorem 58 (**Induced measures).** *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. Suppose $f : \Omega_1 \to \Omega_2$ is ⓜ$\mathcal{F}_1/\mathcal{F}_2$ and μ is a measure on Ω_1. Then the set function defined by*

$$\mu f^{-1}(B) := \mu(f^{-1}(B)), \quad \text{for all } B \in \mathcal{F}_2$$

is a measure on $(\Omega_2, \mathcal{F}_2)$ and is called the **induced measure** *on $(\Omega_2, \mathcal{F}_2)$. Moreover,*

- *μ is a probability measure $\Longrightarrow \mu f^{-1}$ is a probability measure;*
- *μ is a finite measure $\Longrightarrow \mu f^{-1}$ is a finite measure;*
- *μ is a σ-finite measure $\not\Longrightarrow \mu f^{-1}$ is a σ-finite measure.*

Give example of where the induced measure is not σ-finite but the base measure is.

Definition 37 (**Random variable).** *Any function $X\colon \Omega \to \mathbb{R}$ which is measurable $\mathcal{F}/\mathcal{B}^{\mathbb{R}}$ is said to be a* **random variable**.

Distribution function are useful for making random variables with the specified induced distribution.

Definition 38 (**Distribution function on** $\mathbb{R}$**).** *A function $F\colon \mathbb{R} \to \mathbb{R}$ if called a* **distribution function** *if F satisfies the following three requirements:*

- *F is non-decreasing;*
- *F is right continuous;*
- *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.*

Theorem 59 (**Df's determine** PX^{-1}**).** *If F is a distribution function then there exists a random variable X defined on some probability space $(\Omega, \mathcal{F}, P)$ such that*

$$P(X \le x) = F(x) \text{ for all } x \in \mathbb{R}.$$

Moreover, the distribution of X is uniquely determined by F.

Theorem 60 ($F^{-1}(U) \sim X$**).** *Let X be a random variable and define $F(x) := P(X \le x)$. Let U be a random variable uniformly distributed over $(0, 1)$. Then F is a distribution function and $F^{-1}(U) \sim X$ (i.e. $F^{-1}(U)$ and X have the same induced distribution on $\mathbb{R}$) where*

$$F^{-1}(u) := \inf\{x \in \mathbb{R}\colon u \le F(x)\}. \tag{34}$$

Theorem 61 ($F(X) \sim U$**).** *Let X be a random variable with distribution function $F(x) = P(X \le x)$. Then*

$$P(F(X) \le u) \le u \text{ for all } 0 < u < 1.$$

Moreover, F is continuous if and only if $P(F(X) \le u) = u$ for all $0 < u < 1$.

6 σ-fields generated by functions

The results here will be used often in the later text. We will use them for generating a product measure, for Fubini's theorem and for conditional expected value.

Definition 39 (**The σ-field generated by functions).** *Let $\mathcal{I}$ be an arbitrary index set. Let $(\Omega_i, \mathcal{F}_i)$ be a collection of measurable spaces indexed by $i \in \mathcal{I}$. Let $f_i : \Omega \to \Omega_i$ be a collection of functions indexed by $i \in \mathcal{I}$. Then the σ-**field generated by** $\{f_i : i \in \mathcal{I}\}$ is defined as*

$$\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle := \bigcap_{\substack{\mathcal{F} \text{ is a } \sigma\text{-field on } \Omega \\ \text{each } X_i \text{ is } ⓜ \mathcal{F}/\mathcal{F}_i}} \mathcal{F}$$

and corresponds to the smallest σ-field on Ω which makes all the random variables f_i measurable.

When $\mathcal{F}_i$ are clear from context we may, and do, write

$$\sigma\langle f_i : i \in \mathcal{I}\rangle \text{ as shorthand for } \sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle.$$

Theorem 62 (**Pull back for one map).** *For a single function $f_1 : \Omega \to \Omega_1$ where $(\Omega_1, \mathcal{F}_1)$ is a measurable space one has that $\sigma\langle f_1, \mathcal{F}_1\rangle = f^{-1}(\mathcal{F}_1) := \{f^{-1}(F) : F \in \mathcal{F}_1\}$.*

Proof. We immediately have that $f^{-1}(\mathcal{F}_1) \subset \sigma\langle f_1, \mathcal{F}_1\rangle$ since by definition f is ⓜ $\sigma\langle f_1, \mathcal{F}_1\rangle/\mathcal{F}_1$. To show $\sigma\langle f_1, \mathcal{F}_1\rangle \subset f^{-1}(\mathcal{F}_1)$, all we need is to establish that $f^{-1}(\mathcal{F}_1)$ is a σ-field (since trivially f is ⓜ $f^{-1}(\mathcal{F}_1)/\mathcal{F}_1$). This is easily checked by the properties of pull-back sets given in Theorem 50. □

Theorem 63 (**Generators are enough).** *If, in addition to the assumptions presented in Definition 39, one has that each $\mathcal{F}_i = \sigma\langle \mathcal{A}_i\rangle$, then*

$$\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle = \sigma\big\langle f_i^{-1}(A_i) : A_i \in \mathcal{A}_i, i \in \mathcal{I}\big\rangle.$$

Proof. The only interesting direction is to show $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle \subset \sigma\big\langle f_i^{-1}(A_i) : A_i \in \mathcal{A}_i, i \in \mathcal{I}\big\rangle$. By "good sets" we just need to show that each f_i is ⓜ $\sigma\big\langle f_i^{-1}(A_i) : A_i \in \mathcal{A}_i, i \in \mathcal{I}\big\rangle/\sigma\langle \mathcal{A}_i\rangle$. This follows immediately from Theorem 51 (generators are enough). □

Definition 40 (**The product σ-field).** *Let $\mathcal{I}$ be an arbitrary index set. Let $(\Omega_i, \mathcal{F}_i)$ be a collection of measurable spaces indexed by $i \in \mathcal{I}$. Define the product σ-field on $\prod_{i\in\mathcal{I}} \Omega_i$ to be*

$$\textstyle\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i := \sigma\langle \pi_i, \mathcal{F}_i : i \in \mathcal{I}\rangle$$

where $\pi_i : \Omega \to \Omega_i$ is defined as the i^{th} coordinate mapping (e.g. $\pi_i(\omega_1, \omega_2, \ldots) = \omega_i$).

Theorem 64 (**Clump f_i into a vector map).** *Let $(\Omega, \mathcal{F})$ be a measurable space. Let $(\Omega_i, \mathcal{F}_i)$ be a collection of measurable spaces indexed by an arbitrary index set $\mathcal{I}$ and let $f_i : \Omega \to \Omega_i$. Define $\vec{f} : \Omega \to \prod_{i\in\mathcal{I}} \Omega_i$ to be the map which sends $\omega \overset{\vec{f}}{\mapsto} (f_i(\omega))_{i\in\mathcal{I}}$. Then*

$$\vec{f} \text{ is } ⓜ \mathcal{F}/\textstyle\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i \iff \text{each } f_i \, ⓜ \, \mathcal{F}/\mathcal{F}_i.$$

Moreover, $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle = \vec{f}^{-1}\big(\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\big)$.

Theorem 65 (**Measurable with respect to $\sigma\langle f_i\rangle$).** *Using the same assumptions and notation as in Definition 39, a function $f : \Omega \to \bar{\mathbb{R}}$ is measurable $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle$ if and only if there exists a function $g : \prod_{i\in\mathcal{I}} \Omega_i \to \bar{\mathbb{R}}$ which is measurable $\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$ and $f = g((f_i)_{i\in\mathcal{I}})$.*

Proof. ($\Longleftarrow$) This follows directly from Theorem 64 (clump theorem) and Theorem 52 (composition of measurable is measurable).

($\Longrightarrow$) This follow by an application of the $1-2-3$ argument. In particular, we show the result for simple function, then extend by taking point-wise limits. To start let $\vec{f} := (f_i)_{i\in\mathcal{I}}$ denote the clumped vector map. To summarize what is know from the assumptions and 64

- $f : \Omega \to \bar{\mathbb{R}}$ is ⓜ $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle/\mathcal{B}$;
- $\vec{f} : \Omega \to \prod_{i\in\mathcal{I}} \Omega_i$ is ⓜ $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle/\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$.
- $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle = \vec{f}^{-1}\big(\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\big)$

Suppose that $f = \sum_{k=1}^n c_k I_{A_k}$ for $A_k \in \sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle = \vec{f}^{-1}\big(\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\big)$. Since $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle = \vec{f}^{-1}\big(\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\big)$ we can write $A_k = \vec{f}^{-1}(F_k)$ where $F_k \in \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$. Now

$$f = \sum_{k=1}^n c_k I_{A_k} = \sum_{k=1}^n c_k I_{\vec{f}^{-1}(F_k)} = \sum_{k=1}^n c_k I_{F_k} \circ \vec{f} = g \circ \vec{f}$$

where $g := \sum_{k=1}^n c_k I_{F_k}$. Certainly g is ⓜ $\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i/\mathcal{B}$ as was to be shown.

To finish let f be a arbitrary ⓜ $\sigma\langle f_i, \mathcal{F}_i : i \in \mathcal{I}\rangle/\mathcal{B}$ function. By Theorem 57 (the structure theorem) we can write $f(\omega) = \lim f_n(\omega)$ where f_n are bounded simple functions. Therefore, from the preivous case, there exists g_n which are ⓜ $\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i/\mathcal{B}$ and $f_n(\omega) = g_n(\vec{f}(\omega))$. We definitely have that $f(\omega) = \lim f_n(\omega) = \lim g_n(\vec{f}(\omega))$ at each ω. However we are not exactly done since there is no reason that $\lim_n g_n(v)$ exists for v's which are not of the form $v = \vec{f}(\omega)$ (and therefore setting $g(v) := \lim g_n(v)$ only defines g on the range of $\vec{f}$). To get around this set

$$g(v) := \begin{cases} \limsup_n g_n(v), & \text{when } \limsup_n g_n(v) = \liminf_n g_n(v) \\ 0, & \text{otherwise.} \end{cases}$$

This g definitely satisfies $f = g \circ \vec{f}$ and it is measurable since $\limsup_n g_n(v)$ is measurable (since the g_n's are and the closure properties of measurable functions) and the event $\{v : \limsup_n g_n(v) = \liminf_n g_n(v)\}$ is also measurable. □

The following is a corollary of Theorem 65 which will be important when we define conditional expected value. In particular $E(X|Y_1, \ldots, Y_n)$ will be required to be measurable with respect to $\sigma\langle Y_1, \ldots, Y_n\rangle$. The following corollary says that $E(X|Y_1, \ldots, Y_n)$ must then be of the form $g(Y_1, \ldots, Y_n)$ where g is Borel measurable.

Corollary 7. *Let $X, Y_1, \ldots, Y_n$ be functions which map Ω into $\bar{\mathbb{R}}$. Then*

$$X \text{ is } ⓜ\, \sigma\langle Y_1, \ldots, Y_n\rangle \Longleftrightarrow X = g(Y_1, \ldots, Y_n) \text{ where } g \text{ is } ⓜ.$$

Exercise 28. *Show that $\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$ equals $\sigma\langle \Pi_{h\in\mathcal{H}} B_h \colon B_h \in \mathcal{F}_h, \text{countable } \mathcal{H} \subset \mathcal{I}\rangle$.*

Exercise 29. *Show that $\mathcal{B}^{\mathbb{R}^d} = \bigotimes_{i=1}^d \mathcal{B}^{\mathbb{R}}$ and $\mathcal{B}^{\bar{\mathbb{R}}^d} = \bigotimes_{i=1}^d \mathcal{B}^{\bar{\mathbb{R}}}$*

6.1 Application: random variable independence

Section Assumption. *For the rest of this section let $(\Omega, \mathcal{F}, P)$ denote a probability space.*

Definition 41 (Independence for random variables). *A collection of random variables $\{X_i : i \in \mathcal{I}\}$ are said to be* **independent** *if and only if the collection of σ-fields $\{\sigma\langle X_i\rangle \colon i \in \mathcal{I}\}$ are independent.*

Theorem 66 (ANOVA for random variables). *Consider the following array of random variables all defined on the same probability space $(\Omega, \mathcal{F}, P)$*

$$\begin{array}{ccc} X_{1,1} & X_{1,2} & \cdots \\ X_{2,1} & X_{2,2} & \cdots \\ X_{3,1} & X_{3,2} & \cdots \\ \vdots & \vdots & \ddots \end{array}$$

Each row may have a different number of columns (finite or infinite) and the number of rows may be finite or infinite. Let $\mathscr{R}_1, \mathscr{R}_2, \ldots$ denote the σ-fields generated by the rows: $\mathscr{R}_i := \sigma\langle X_{i,1}, X_{i,2}, \cdots\rangle$. Then the full collection $\{X_{i,k}\}$ of random variables are independent if and only if the following two statements hold:

1. *The random variables within each row are independent;*
2. *The σ-fields generated by the rows, $\mathscr{R}_1, \mathscr{R}_2, \ldots$, are independent.*

Theorem 67 (Existence of independent $X_1, X_2, \ldots$). *Let $\mu_1, \mu_2, \ldots$ be a finite or infinite sequence of probability measures on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$. Then there exists on some probability space $(\Omega, \mathcal{F}, P)$ a sequence of independent random variables $X_1, X_2, \ldots$ such that X_i has distribution μ_i for each i.*

Theorem 68 (Kolmogorov's 0-1 law for random variables). *Let $X_1, X_2, \ldots$ be an infinite sequence of independent random variables on a probability space $(\Omega, \mathcal{F}, P)$. Then all tail events in the tail σ-field $\mathcal{T} := \bigcap_{n=1}^\infty \sigma\langle X_n, X_{n+1}, \ldots\rangle$ have probability either 0 or 1 and all functions $f : \Omega \to \bar{\mathbb{R}}$ which are ⓜ $\mathcal{T}/\mathcal{B}$ are almost surely constant.*

Definition 42 (Symmetric function). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. Another random variable Y on $(\Omega, \mathcal{F}, P)$ is said to be a* **symmetric function** *of the X_n's if $Y = f(X_1, X_2, \ldots)$ where $f \colon \mathbb{R}^\infty \to \mathbb{R}$ is ⓜ $\bigotimes_{i=1}^\infty \mathcal{B}^{\mathbb{R}}/\mathcal{B}^{\mathbb{R}}$ and $f(x_1, x_2, \ldots) = f(x_{\pi_1}, x_{\pi_2}, \ldots)$ whenever π is a permutation that permutes a finite number coordinates. We say an event $A \in \mathcal{F}$* **depends symmetrically** *on the X_n's if the indicator function $I_A(w)$, for $w \in \Omega$, is a symmetric function of the X_n's.*

Theorem 69 (Hewitt-Savage 0-1 law). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. Each event that depends symmetrically on the X_n's has probability 0 or 1, and each random variable that is a symmetric function of the X_n's is almost surely constant*

Exercise 30. *Suppose that $Y_1, Y_2, \ldots$ is an infinite sequence of independent random variables, all defined on the same probability space $(\Omega, \mathcal{F}, P)$, taking the values 0 and 1 with probability 1/2 each. Show that $U := \sum_{k=1}^\infty 2^{-k} Y_k$ is uniformly distributed on $[0, 1]$. Hint: show*

$$P[U \le x] = \begin{cases} x & \text{when } x \in [0,1]; \\ 1 & \text{when } x > 1; \\ 0 & \text{when } x < 0. \end{cases}$$

for all $x \in \mathbb{R}$ by analyzing $P[U_n \le x]$ as $n \to \infty$ where $U_n := \sum_{k=1}^n 2^{-k} Y_k$.

Suppose X and Y are two random variables, not necessarily defined on the same probability space. Y is said to be **stochastically larger** than X if $P[X \le x] \ge P[Y \le x]$ for all $x \in \mathbb{R}$.

Exercise 31. *Suppose X and Y are random variables and that Y is stochastically larger than X. Show there exists random variables X^* and Y^* defined on a common probability space $(\Omega, \mathcal{F}, P)$ such that $X^* \sim X$, $Y^* \sim Y$ and $X^*(\omega) \le Y^*(\omega)$ for all $\omega \in \Omega$.*

Let $\mathcal{I}$ be an arbitrary index set and let $X_i, i \in \mathcal{I}$ be a family of random variables where each X_i is defined on a probability space $(\Omega_i, \mathcal{F}_i, P_i)$. Let $F_i(x) := P_i(X_i \le x)$ be the distribution function of X_i. The X_i's are said to be **stochastically dominated** by a random variable X if X is stochastically larger than $|X_i|$ for each $i \in \mathcal{I}$. The X_i's are said to be **pointwise dominated** by X if all the random variables X, X_i, for $i \in \mathcal{I}$, are defined on the same probability space and $|X_i(\omega)| \le X(\omega)$ for each $w \in \Omega$ and for each $i \in \mathcal{I}$. The distribution functions

F_i are said to be **tight** if the following two equalities hold

$$\lim_{x\to-\infty}\sup_{i\in\mathcal{I}} F_i(x) = 0$$
$$\lim_{y\to+\infty}\inf_{i\in\mathcal{I}} F_i(y) = 1.$$

Exercise 32. *Let $X_i, i \in \mathcal{I}$, be a family of random variables where each X_i is defined on a probability space $(\Omega_i, \mathcal{F}_i, P_i)$. Show that the following are equivalent*

1. *The X_i's are stochastically dominated by some random variable;*

2. *The corresponding distribution functions F_i are tight;*

3. *There exists random variables X_i^*, $i \in \mathcal{I}$, all defined on a common probability space such that $X_i^* \sim X_i$ for each $i \in \mathcal{I}$ and the X_i^*'s are pointwise dominated by some random variable.*

7 Construction of $\int_\Omega f d\mu$

Definition 43 (Definition of $\int_\Omega f d\mu$ for $f \in \mathscr{N}_s$). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \bar{\mathbb{R}}$ be in $\mathscr{N}_s$. Then $f = \sum_{i=1}^n c_i I_{A_i}$ for some $c_1, \ldots, c_n \in [0, \infty]$ and disjoint $\mathcal{F}$-sets $A_1, \ldots, A_n$. The integral of f with respect to μ is defined as*

$$\int_\Omega f d\mu := \sum_{i=1}^n c_i \mu(A_i).$$

Theorem 70 (The simple 3). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

1. *If $f \in \mathscr{N}_s$ then $\int_\Omega f d\mu$ is well defined.*
2. **Monotonicity:** *If $f, g \in \mathscr{N}_s$ then*
$$f \leq g \Longrightarrow \int_\Omega f d\mu \leq \int_\Omega g d\mu.$$
3. **Linearity:** *If both $f, g \in \mathscr{N}_s$ then*
$$\int_\Omega \alpha f + \beta g d\mu = \alpha \int_\Omega f d\mu + \beta \int_\Omega g d\mu \tag{35}$$
for all $\alpha, \beta \in [0, \infty]$.
4. **Continuous from below:** *If $f_1, f_2, \ldots$ and f are in $\mathscr{N}_s$ then*
$$f_n \uparrow f \Longrightarrow \int_\Omega f_n d\mu \uparrow \int_\Omega f d\mu.$$

Definition 44 (Definition of $\int_\Omega f d\mu$ for $f \in \mathscr{N}$). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \bar{\mathbb{R}}$ be in $\mathscr{N}$. Then there exists $f_n \in \mathscr{N}_s$ such that $f_n \uparrow f$. The integral of f with respect to μ is defined as*

$$\int_\Omega f d\mu := \lim_{n \to \infty} \int_\Omega f_n d\mu.$$

Theorem 71 (The little 3). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

1. *If $f \in \mathscr{N}$ then $\int_\Omega f d\mu$ is well defined.*
2. **Monotonicity:** *If $f, g \in \mathscr{N}$ then*
$$f \leq g \Longrightarrow \int_\Omega f d\mu \leq \int_\Omega g d\mu.$$
3. **Linearity:** *If both $f, g \in \mathscr{N}$ then*
$$\int_\Omega \alpha f + \beta g d\mu = \alpha \int_\Omega f d\mu + \beta \int_\Omega g d\mu \tag{36}$$
for all $\alpha, \beta \in [0, \infty]$.
4. **Continuous from below:** *If $f_1, f_2, \ldots$ and f are in $\mathscr{N}$ then*
$$f_n \uparrow f \Longrightarrow \int_\Omega f_n d\mu \uparrow \int_\Omega f d\mu.$$

Theorem 72 (Useful side facts). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

1. *If $f \in \mathscr{N}$ and $\int_\Omega f d\mu < \infty$ then $f < \infty$ μ-a.e..*
2. *If $f \in \mathscr{N}$ then*
$$\int_\Omega f d\mu = 0 \Longleftrightarrow f = 0 \ \mu\text{-a.e.}.$$
3. *If $f, g \in \mathscr{N}$ and $f = g$ μ-a.e. then $\int_\Omega f d\mu = \int_\Omega g d\mu$.*

Definition 45 (Extending $\int_\Omega f d\mu$ to some–but not all–ⓜ$\mathcal{F}$ functions). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \bar{\mathbb{R}}$ be an $\mathcal{F}$ measurable function such that either $\int_\Omega f^+ d\mu < \infty$ or $\int_\Omega f^- d\mu < \infty$. The integral of f with respect to μ is defined as*

$$\int_\Omega f d\mu := \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu.$$

Remark. *One consequence of Theorem 72 (ii) is that for any function $f \colon \Omega \to \bar{\mathbb{R}}$, such that $\int f d\mu$ is defined, we are free to change the value of $f(w)$ on a μ-negligible set without changing the value of the integral so long as the new function is still measurable.*

Definition 46 (Quasi-integrable and integrable). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

- *$Q^+(\mu)$ denotes the set of functions $f : \Omega \to \bar{\mathbb{R}}$ which are measurable $\mathcal{F}$ and $\int_\Omega f^+ d\mu < \infty$*
  (use $Q^+(\Omega, \mathcal{F}, \mu)$ if we want to be specific about Ω and $\mathcal{F}$);
- *$Q^-(\mu)$ denotes the set of functions $f : \Omega \to \bar{\mathbb{R}}$ which are measurable $\mathcal{F}$ and $\int_\Omega f^- d\mu < \infty$;*
- *$Q(\mu) := Q^+(\mu) \cup Q^-(\mu)$;*
- *$L_1(\mu) := Q^+(\mu) \cap Q^-(\mu)$.*

Definition 47 (Extending $\int_\Omega f d\mu$ to some–but not all—functions only defined μ-a.e.). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \cap A \to \bar{\mathbb{R}}$ where A^c is a μ-null set (i.e. f is defined μ-a.e.). If it is possible to change or define f on a μ-null cover of A^c, to yield a function $f^* : \Omega \to \bar{\mathbb{R}}$ which is $f \in Q(\mu)$, then we define*

$$\int_\Omega f d\mu := \int_\Omega f^* d\mu.$$

Remark. *The above definition is useful for the next theorem since it allows us to potentially integrate functions such as $f + g$ even when there is a μ-null set of w's such that $f(w) + g(w) = \infty - \infty$.*

7.1 The Big Three: monotonicity, linearity and continuity from below

Theorem 73 (The Big Three). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

1. **Monotonicity:** *If $f, g \in Q(\mu)$ then*
$$f \leq g \ \mu\text{-a.e.} \Longrightarrow \int_\Omega f d\mu \leq \int_\Omega g d\mu.$$
2. **Linearity:**
If $f \in Q(\mu)$ and $\alpha \in \mathbb{R}$ then $\alpha f \in Q(\mu)$ and
$$\int_\Omega \alpha f \, d\mu = \alpha \int_\Omega f \, d\mu.$$
If $f \in \mathscr{N}$ and $\alpha \in \{-\infty, \infty\}$ then $\alpha f \in Q(\mu)$ and
$$\int_\Omega \alpha f \, d\mu = \alpha \int_\Omega f \, d\mu.$$
If f and g are such that the sum $\int_\Omega f \, d\mu + \int_\Omega g \, d\mu$ is defined (if both $f, g \in Q^+(\mu)$ or if both $f, g \in Q^-(\mu)$) then $f + g \in Q(\mu)$ and
$$\int_\Omega f + g \, d\mu = \int_\Omega f \, d\mu + \int_\Omega g \, d\mu.$$
3. **Continuous from below:** *If $f_1, f_2, \ldots$ are measurable $\mathcal{F}$ then*
$$0 \leq f_n \uparrow f \ \mu\text{-a.e.} \Longrightarrow \int_\Omega f_n d\mu \uparrow \int_\Omega f d\mu.$$

Corollary 8 (Facts embedded in the proof of Big 3).

- *If $g \in Q^+(\mu)$, f is ⓜ$\mathcal{F}$ and $f \leq g$ a.e. then $f \in Q^+(\mu)$;*
- *If $f \in Q^-(\mu)$, g is ⓜ$\mathcal{F}$ and $f \leq g$ a.e. then $g \in Q^-(\mu)$;*
- *If $f \in Q^\pm(\mu)$ and $\alpha \in [0, \infty)$ then $\alpha f \in Q^\pm(\mu)$;*
- *If $f \in Q^\pm(\mu)$ and $\alpha \in (-\infty, 0)$ then $\alpha f \in Q^\mp(\mu)$;*
- *If $f, g \in Q^\pm(\mu)$ then $f + g \in Q^\pm(\mu)$;*

Corollary 9. *Suppose $f, g \in Q(\mu)$ and either $f \in L_1(\mu)$ or $g \in L_1(\mu)$. Then if $\alpha, \beta \in \mathbb{R}$ one has that $\alpha f + \beta g \in Q(\mu)$ and*
$$\int_\Omega \alpha f + \beta g \, d\mu = \alpha \int_\Omega f \, d\mu + \beta \int_\Omega g \, d\mu.$$

Corollary 10.

- *$\left|\int f d\mu\right| \leq \int |f| d\mu$ for all $f \in Q(\mu)$;*
- *If f is ⓜ$\mathcal{F}$ and $\int |f| d\mu < \infty$ then $f \in L_1(\mu)$.*

Exercise 33 (Relate with Billingsley's definition of $\int_\Omega f d\mu$). *Show that for any $f \in \mathscr{N}$*
$$\int_\Omega f d\mu = \sup \left\{ \sum_i \left[\mu(A_i) \inf_{w \in A_i} f(w)\right] : \{A_i\} \in \mathcal{A} \right\}$$
where $\mathcal{A}$ is the collection of finite $\mathcal{F}$-partitions $\{A_i\}$ of Ω.

Exercise 34 (More general continuity results for $\int_\Omega f d\mu$). *Suppose $f_1, f_2, \ldots$ are measurable $\mathcal{F}$ functions on Ω. Show the following two statements:*

1. *If $f_1 \in Q^-(\mu)$ and $f_n \uparrow f$ then $f_n \in Q^-(\mu)$ for all $n \geq 1$, $f \in Q^-(\mu)$ and $\int_\Omega f_n \, d\mu \uparrow \int_\Omega f \, d\mu$.*
2. *If $f_1 \in Q^+(\mu)$ and $f_n \downarrow f$ then $f_n \in Q^+(\mu)$ for all $n \geq 1$, $f \in Q^+(\mu)$ and $\int_\Omega f_n \, d\mu \downarrow \int_\Omega f \, d\mu$.*

Exercise 35 (Piecewise monotonicity). *Let $f_1, f_2, \ldots$ be a sequence of measurable $\mathcal{F}$ functions on Ω. Show that if:*

- *there exists an $\mathcal{F}$-set $B \subset \Omega$;*
- *$f_n(w) \uparrow f(w)$ for each $w \in B$;*
- *$f_n(w) \downarrow f(w)$ for each $w \in B^c$;*
- *$f_1 \in L_1(\mu)$ and $f \in Q(\mu)$,*

then $f_n \in Q(\mu)$ for all n and $\int f_n \, d\mu \to \int f \, d\mu$.

8 Change of variables and densities

Theorem 74 (**Change of variables).** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $(\Omega', \mathcal{F}')$ be measurable space. Suppose $\Omega \xrightarrow{T} \Omega' \xrightarrow{f} \bar{\mathbb{R}}$ where T is measurable $\mathcal{F}/\mathcal{F}'$ and f is measurable $\mathcal{F}'/\mathcal{B}$. Then $f \in Q^{\pm}(\Omega', \mathcal{F}', \mu T^{-1})$ if and only if $f \circ T \in Q^{\pm}(\Omega, \mathcal{F}, \mu)$ and either one imply that*

$$\int_{\Omega} f \circ T d\mu = \int_{\Omega'} f d\mu T^{-1}.$$

Definition 48 (**Indefinite integral).** *If $f \in Q(\mu)$ then the set function $\int_{\bullet} f d\mu : \mathcal{F} \to \bar{\mathbb{R}}$ defined as*

$$A \mapsto \int_A f d\mu := \int_{\Omega} f I_A d\mu$$

for all $A \in \mathcal{F}$ is called the **indefinite integral** *of f with respect to μ.*

Theorem 75. *If $f \in Q(\mu)$ then $\int_{\bullet} f d\mu$ is countably additive over disjoint sets.*

Proof. Let $F_1, F_2, \ldots$ are disjoint $\mathcal{F}$-sets. We use the 2-3 argument.

(*Step 2, prove for $f \in \mathcal{N}$*)

$$\begin{aligned}
\int_{\cup_k F_k} f d\mu &:= \int_{\Omega} f I_{\cup_k F_k} d\mu \\
&= \int_{\Omega} f \sum_{k=1}^{\infty} I_{F_k} d\mu, \text{ disjoint } F_k \\
&= \int_{\Omega} \sum_{k=1}^{\infty} f I_{F_k} d\mu \\
&= \int_{\Omega} \lim_n\uparrow \sum_{k=1}^{n} f I_{F_k} d\mu, \; f I_{F_k} \geq 0 \\
&= \lim_n\uparrow \int_{\Omega} \sum_{k=1}^{n} f I_{F_k} d\mu \text{ by Big 3 and } \sum_{k=1}^{n} f I_{F_k} \geq 0 \\
&= \lim_n\uparrow \sum_{k=1}^{n} \int_{\Omega} f I_{F_k} d\mu \text{ by Big 3 and } f I_{F_k} \geq 0 \\
&= \sum_{k=1}^{\infty} \int_{F_k} f d\mu
\end{aligned}$$

(*Step 3, prove for $f \in Q(\mu)$*)

$$\int_{\cup_k F_k} f d\mu = \int_{\cup_k F_k} f^+ d\mu - \int_{\cup_k F_k} f^- d\mu$$

At least one term above is finite by $f \in Q(\mu)$

$$\begin{aligned}
&= \sum_{k=1}^{\infty} \int_{F_k} f^+ d\mu - \sum_{k=1}^{\infty} \int_{F_k} f^- d\mu, \text{ by Step 2} \\
&= \sum_{k=1}^{\infty} \Big[\int_{F_k} f^+ d\mu - \int_{F_k} f^- d\mu \Big]
\end{aligned}$$

By Corollary 9 with counting measure since both sequences $\{\int_{F_k} f^{\pm} d\mu\}_{k=1}^{\infty}$ are in $Q^-(\#)$ and at least one is in $L_1(\#)$.

$$= \sum_{k=1}^{\infty} \int_{F_k} f d\mu$$

□

Corollary 11 (**Indefinite integrals are measures).** *If $f \in \mathcal{N}$ then $\int_{\bullet} f d\mu$ is a measure. If, in addition, $\int_{\Omega} f\, d\mu = 1$ then $\int_{\bullet} f d\mu$ is a probability measure.*

Definition 49 (**Densities).** *For any measure ν on the measurable space $(\Omega, \mathcal{F})$, if there exists $\delta \in \mathcal{N}$ such that $\nu(\bullet) = \int_{\bullet} \delta d\mu$ over $\mathcal{F}$ then we say that δ* **is the density of ν with respect to μ**.

Theorem 76. *Let $f, g \in Q(\mu)$. If $f \in L_1(\mu)$ or $g \in L_1(\mu)$ or μ is σ-finite then*

$$\int_{\bullet} f d\mu \leq \int_{\bullet} g d\mu \text{ on } \mathcal{F} \iff f \leq g \;\; \mu\text{-a.e.}$$

It is clear that the above theorem does not hold without some condition like $f \in L_1(\mu)$ or $g \in L_1(\mu)$ or μ is σ-finite. Indeed a counter example can be found by

$$\begin{aligned}
\Omega &:= \mathbb{R} \\
\mathcal{F} &:= \{\varnothing, \Omega, (-\infty, 0), [0, \infty)\} \\
\mu &:= \mathcal{L}_1 \text{ on } \mathcal{F} \\
f &:= 2 \\
g &:= 1.
\end{aligned}$$

Now clearly $f \not\leq g$ but $\int_{\bullet} f d\mu \leq \int_{\bullet} g d\mu$ on $\mathcal{F}$.

Proof. The direction ($\Longleftarrow$) follows directly by monotonicity in Big 3. To show ($\Longrightarrow$) assume $\int_{\bullet} f d\mu \leq \int_{\bullet} g d\mu$ on $\mathcal{F}$. We show $\mu(f > g) = 0$.

•(*Case 1:* $f \in L_1(\mu)$ or $g \in L_1(\mu)$)

$$\begin{aligned}
& f I_{\{f>g\}} \geq g I_{\{f>g\}} \\
&\Longrightarrow \int_{\{f>g\}} f d\mu \geq \int_{\{f>g\}} g d\mu, \text{ Big 3} \\
&\Longrightarrow \int_{\{f>g\}} f d\mu = \int_{\{f>g\}} g d\mu, \text{ since } \int_{\bullet} f d\mu \leq \int_{\bullet} g d\mu \\
&\Longrightarrow \int f I_{\{f>g\}} d\mu = \int g I_{\{f>g\}} d\mu \\
&\Longrightarrow \int \underbrace{(f-g) I_{\{f>g\}}}_{\in \mathcal{N}} d\mu = 0, \text{ since } f \in L_1(\mu) \text{ or } g \in L_1(\mu) \\
&\Longrightarrow (f-g) I_{\{f>g\}} = 0 \;\; \mu\text{-a.e. by Useful facts} \\
&\Longrightarrow \mu(\{f > g\}) = 0, \text{ since } (f-g) > 0 \text{ when } I_{\{f>g\}} = 1
\end{aligned}$$

•(*Case 2: μ is finite*) Let $A_n := \{|f| < n\}$. Now since μ is a finite measure $fI_{A_n} \in L_1(\mu)$ and $gI_{A_n} \in Q(\mu)$. Moreover

$$\int_\bullet fI_{A_n} d\mu = \int_{\bullet\cap A_n} f d\mu \le \int_{\bullet\cap A_n} g d\mu = \int_\bullet gI_{A_n} d\mu$$

on $\mathcal{F}$. Therefore by case 1, $fI_{A_n} \le gI_{A_n}$ μ-a.e. for every n. This gives

$$f \le g \ \mu\text{-a.e on} \bigcup_{n=1}^{\infty} A_n = \{|f| < \infty\}. \tag{37}$$

Similary we can define $B_n := \{|g| < n\}$ and using the same argument as above conclude that

$$f \le g \ \mu\text{-a.e on} \bigcup_{n=1}^{\infty} B_n = \{|g| < \infty\}. \tag{38}$$

We also have that

$$\begin{align}
&f \le g \text{ on } \{f = \infty\} \cap \{g = \infty\} \tag{39}\\
&f \le g \text{ on } \{f = -\infty\} \cap \{g = \infty\} \tag{40}\\
&f \le g \text{ on } \{f = -\infty\} \cap \{g = -\infty\} \tag{41}\\
& \tag{42}
\end{align}$$

The last case $\{f = \infty\} \cap \{g = -\infty\}$ must have μ-measure zeros or else it would contradict $\int_\bullet f d\mu \le \int_\bullet g d\mu$. Considering the union of all the sets in (37)-(41) gives

$$f \le g \ \mu\text{-a.e.}$$

as was to be shown.

•(*Case 3: μ is σ-finite*) Let $\Omega = \bigcup_{n=1}^{\infty} F_n$ where F_n are disjoint $\mathcal{F}$-sets having finite μ-measure. Notice that

$$\mu(f > g) = \sum_{n=1}^{\infty} \underbrace{\mu(\{f > g\} \cap F_n)}_{=:\mu_n(f>g)}$$

where μ_n is a finite measure. Case 2 now implies $\mu_n(f > g) = 0$ since

$$\begin{aligned}
\int_\bullet f d\mu_n &= \int_\bullet fI_{F_n} d\mu \ \text{ by 1-2-3 argument}\\
&= \int_{\bullet\cap F_n} f d\mu\\
&\le \int_{\bullet\cap F_n} g d\mu, \text{ by assumption}\\
&= \int_\bullet g d\mu_n \ \text{ by 1-2-3 argument}
\end{aligned}$$

□

Corollary 12 (Uniqueness of densities). *Let $f, g \in Q(\mu)$. If $f \in L_1(\mu)$ or $g \in L_1(\mu)$ or μ is σ-finite then*

$$\int_\bullet f d\mu = \int_\bullet g d\mu \text{ on } \mathcal{F} \iff f = g \ \mu\text{-a.e.}$$

Corollary 13 (Uniqueness of densities for probabilities). *The density of any finite measure is unique.*

The next theorem tells us how to compute $\int_\Omega f d\nu$ when $\nu(\bullet) = \int_\bullet \delta d\mu$ for some density δ.

Theorem 77 (Slap in the density: $d\nu = \delta d\mu$). *Let ν and μ be measures on the measurable space $(\Omega, \mathcal{F})$. Suppose ν has density δ with respect to μ. Then $f \in Q^\pm(\nu)$ if and only if $f\delta \in Q^\pm(\mu)$ and either one implies*

$$\int_\Omega f d\nu = \int_\Omega f\delta d\mu. \tag{43}$$

Proof. We use the 1-2-3 argument.

•(*Step 1: Show (43) for $f \in \mathscr{N}_s$*) By non-negativity and closure theorem for ⓜ both f and $f\delta$ are quasi-integrable from below. Therefore

$$\begin{aligned}
\int_\Omega f d\nu &= \int_\Omega \sum_{i=1}^{n} c_i I_{A_i} d\nu, \text{ by Structure Thm}\\
&= \sum_{i=1}^{n} c_i \nu(A_i), \text{ definition of } \int_\Omega\\
&= \sum_{i=1}^{n} c_i \int_\Omega \delta I_{A_i} d\mu\\
&= \int_\Omega \delta \sum_{i=1}^{n} c_i I_{A_i} d\mu, \text{ by Little 3}\\
&= \int_\Omega \delta f d\mu.
\end{aligned}$$

Notice the integrals in the above equality could all be ∞.

•(*Step 2: Show (43) for $f \in \mathscr{N}$*) The follows directly by monotonicity in Little 3.

•(*Step 3: Show the whole theorem for general f*) From step 2 we have

$$\int_\Omega f^\pm d\nu = \int f^\pm \delta d\mu = \int (f\delta)^\pm d\mu.$$

Therefore $f \in Q^\pm(\nu) \iff f\delta \in Q^\pm(\mu)$ and either implies (43).

□

The previous theorem gives more motivation for the notation that a density δ of ν wrt μ should be written $\frac{d\nu}{d\mu}$. Indeed "slap in the density" says $d\nu = \delta d\mu$. At times, I will write $d\nu = \delta d\mu$ as short hand for the statement that ν has a density δ with respect to μ. I will also, at times, say that $\frac{d\nu}{d\mu}$ exists, by which I mean that there exists a density $\frac{d\nu}{d\mu}$ of ν with respect to μ. Note that $\frac{d\nu}{d\mu}$ is unique when either μ is σ-finite or $\frac{d\nu}{d\mu} \in L_1(\mu)$.

Theorem 78 (**The chain rule for densities).** *Let ρ, ν and μ be measures on the measurable space $(\Omega, \mathcal{F})$ such that μ is σ-finite. If $\frac{d\rho}{d\nu}$ is a density of ρ with respect to ν and $\frac{d\nu}{d\mu}$ is the density of ν with respect to μ, then $\frac{d\rho}{d\mu}$ exists and*

$$\frac{d\rho}{d\mu} = \frac{d\rho}{d\nu}\frac{d\nu}{d\mu}, \ \mu\text{-a.e.}$$

Proof. We simply need to check that $\int_\bullet \frac{d\rho}{d\nu}\frac{d\nu}{d\mu}d\mu$ gives $\rho(\bullet)$ and then the σ-finite assumption tells us that it is μ-a.e. unique. Indeed, for any $A \in \mathcal{F}$

$$\begin{aligned}\int_A \frac{d\rho}{d\nu}\frac{d\nu}{d\mu}d\mu &= \int_A \frac{d\rho}{d\nu}d\nu, \ \text{by "slap in the density"} \\ &= \int_A d\rho, \ \text{by "slap in the density"} \\ &= \rho(A).\end{aligned}$$

□

Theorem 79 (**The chain rule for densities***). *Let ρ, ν and μ be measures on the measurable space $(\Omega, \mathcal{F})$. If $\frac{d\rho}{d\nu}$ is a density of ρ with respect to ν and $\frac{d\nu}{d\mu}$ is a density of ν with respect to μ, then $\frac{d\rho}{d\nu}\frac{d\nu}{d\mu}$ serves as a (possibly non-unique) density of ρ with respect to μ.*

Theorem 80 (**Change of variables for densities).** *Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces and let μ and ρ be two measures on $(\Omega, \mathcal{F})$ such that μ is σ-finite. Let $T : \Omega \to \Omega'$ be an invertable map of Ω onto Ω' such that T is measurable $\mathcal{F}/\mathcal{F}'$ and T^{-1} is measurable $\mathcal{F}'/\mathcal{F}$. If ρ has density $\frac{d\rho}{d\mu}$ w.r.t μ then ρT^{-1} has density $\frac{d\rho T^{-1}}{d\mu T^{-1}}$ with respect to μT^{-1} and*

$$\frac{d\rho T^{-1}}{d\mu T^{-1}} = \frac{d\rho}{d\mu} \circ T^{-1}, \ \mu T^{-1}\text{-a.e.}$$

Proof.

•(*Show that $\frac{d\rho}{d\mu} \circ T^{-1}$ serves as a density of ρT^{-1} wrt μT^{-1}*)
Let $A \in \mathcal{F}'$. Clearly $\frac{d\rho}{d\mu} \circ T^{-1} \in Q^-(\mu T^{-1})$ by positivity and the fact that composition of measurable functions is measurable. Now

$$\begin{aligned}\int_A \frac{d\rho}{d\mu} \circ T^{-1} d\mu T^{-1} &= \int_{T^{-1}(A)} \frac{d\rho}{d\mu} \circ T^{-1} \circ T d\mu, \\ &\qquad \text{by change of variable thm} \\ &= \int_{T^{-1}(A)} \frac{d\rho}{d\mu} d\mu \\ &= \rho(T^{-1}(A))\end{aligned}$$

•(*Show that μT^{-1} is σ-finite*) Once we establish this we get uniqueness and hence conclude the proof. Notice this result would not necessarily be true if we did not have the additional assumption on T^{-1}. Let $\Omega = \bigcup_{k=1}^\infty A_k$ be a σ-finite cover wrt μ. We show $\Omega' = \bigcup_{k=1}^\infty T(A_k)$ gives a σ-finite cover wrt μT^{-1}.

- Since T maps onto Ω', $\{T(A_k)\}_{k=1}^\infty$ covers Ω'.
- Notice that $T(A_k) \in \mathcal{F}'$ since $T(A_k) = (T^{-1})^{-1}(A_k)$ and T^{-1} is Ⓜ $\mathcal{F}'/\mathcal{F}$.
- Finally notice that $\mu T^{-1}(T(A_k)) = \mu(T^{-1} \circ T(A_k)) = \mu(A_k) < \infty$.

Therefore μT^{-1} is σ-finite.

□

Theorem 81 (**Change of variables for densities***). *Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces and let μ and ρ be measures on $(\Omega, \mathcal{F})$. Let $T : \Omega \to \Omega'$ be an invertible map of Ω onto Ω' such that T is measurable $\mathcal{F}/\mathcal{F}'$ and T^{-1} is measurable $\mathcal{F}'/\mathcal{F}$. If ρ has a density $\frac{d\rho}{d\mu}$ with respect to μ then $\frac{d\rho}{d\mu} \circ T^{-1}$ serves as a (possibly non-unique) density for ρ with respect to μT^{-1}.*

Theorem 82 (**Probabilist's world view of measure theory).** *If μ is a nontrivial (i.e. $\mu \not\equiv 0$) σ-finite measure on the measurable space $(\Omega, \mathcal{F})$, then there exists a density $\delta : \Omega \to (0, \infty)$ and a probability measure P on $(\Omega, \mathcal{F})$ such that*

$$\mu(A) := \int_A \delta dP$$

for all $A \in \mathcal{F}$ (i.e., $\delta = \frac{d\mu}{dP}$).

Proof. Let $\Omega = \bigcup_{k=1}^\infty A_k$ be a σ-finite partition wrt μ. Additionally suppose $0 < \mu(A_k) < \infty$ for each k by absorbing any A_j such that $\mu(A_j) = 0$ into a A_k with $\mu(A_k) > 0$.

Now set

$$\begin{aligned}\delta^* &:= \sum_{k=1}^\infty \frac{w_k}{\mu(A_k)} I_{A_k} \\ P(\bullet) &:= \int_\bullet \delta^* d\mu\end{aligned}$$

where $w_k > 0$ and $\sum_{k=1}^\infty w_k = 1$. Notice that P is a probability measure since

$$\begin{aligned}P(\Omega) &= \int_\Omega \delta^* d\mu \\ &= \int_\Omega \lim_n\uparrow \sum_{k=1}^n \frac{w_k}{\mu(A_k)} I_{A_k} d\mu \\ &= \lim_n\uparrow \sum_{k=1}^n \int_\Omega \frac{w_k}{\mu(A_k)} I_{A_k} d\mu, \ \text{by L3} \\ &= \sum_{k=1}^\infty w_k = 1.\end{aligned}$$

Now define $\delta := 1/\delta^* = \sum_{k=1}^\infty \frac{\mu(A_k)}{w_k} I_{A_k}$ and notice that

$$\int_A \delta dP \overset{\text{slap}}{=} \int_A \delta\delta^* d\mu = \mu(A).$$

Therefore μ has density δ wrt P. □

Exercise 36 (**When is $\int_\bullet \delta d\mu$ finite or σ-finite?**). *Suppose that ρ is a measure with density δ with respect to a measure μ. Show that:*

1. *ρ is finite if and only if δ is integrable;*
2. *If ρ is σ-finite then $\delta < \infty$ μ-a.e.;*
3. *If $\delta < \infty$ μ-a.e. and μ is σ-finite, then ρ is σ-finite;*
4. *Show by example that the conclusion to 3 may fail if the assumption that μ is σ-finite is dropped.*

Exercise 37 (**Conditions for $d\mu/d\rho = 1/(d\rho/d\mu)$**). *Suppose that ρ is a measure with density δ with respect to μ . Show that:*

1. *μ has density $1/\delta$ with respect to ρ if and only if $0 < \delta < \infty$ μ-a.e.;*
2. *If μ is σ-finite and μ has some density, say f, with respect to ρ then $f = 1/\delta$ ρ-a.e and μ-a.e..*

Hint for 1: first find $\int_\bullet 1/\delta d\rho$.

Exercise 38 (**Approximating functions in $L_1(\mu)$**). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Show the following statements:*

1. *If $f \in L_1(\mu)$ then for each $\epsilon > 0$ there exists an integrable simple function g such that $\int |f - g| d\mu \le \epsilon$;*
2. *If $\mathcal{F}_0$ is a field generating $\mathcal{F}$ and μ is finite on $\mathcal{F}_0$, then the function g from (i) can be taken to be of the form $g = \sum_{k=1}^n c_k I_{A_k}$ where each $A_k \in \mathcal{F}_0$.*
3. *Show by example that the conclusion to part 2 may be false if μ is not σ-finite on $\mathcal{F}_0$.*

Exercise 39. *Suppose $f\colon \mathbb{R} \to \mathbb{R}$ and $f \in L_1(\mathcal{L}^1)$. Show that*

$$\lim_{t\to 0} \int |f(x+t) - f(x)| dx = 0.$$

8.1 Application: random variable expected value and densities

Section Assumption. *For the rest of this Section let $(\Omega, \mathcal{F}, P)$ denote a probability space.*

Definition 50 (**Distribution of X**). *If $X\colon \Omega \to \mathbb{R}$ is a random variable, then the induced probability measure PX^{-1} on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$ is called the* **law** *or* **distribution** *of X and is sometimes denoted $\mathcal{L}_X$.*

Notice that the theory on densities developed above unifies probability density functions and probabilty mass functions for continuous versus discrete random variables. For example a binomial random variable has an induced distribution on $\mathbb{R}$ which has density

$$\delta(x) := \binom{n}{x} p^x (1-p)^x I_{\mathbb{Z}^+}(x)$$

with respect to counting measure on $\mathcal{B}^{\mathbb{R}}$. In particular for any even $B \in \mathcal{B}^{\mathbb{R}}$ and any $X \sim Bin(n,p)$ we have

$$\begin{aligned}
\int_B \delta(x) d\#(x) &= \int \binom{n}{x} p^x (1-p)^x I_{\mathbb{Z}^+}(x) I_B(x) d\#(x) \\
&= \int \sum_{k=0}^n \binom{n}{k} p^k (1-p)^k I_{\{k\}\cap B}(x) d\#(x) \\
&= \sum_{k=0}^n \binom{n}{k} p^k (1-p)^k \# \big[\{k\} \cap B\big] \\
&= P(X \in B) \\
&= PX^{-1}(B)
\end{aligned}$$

A probability density function for a continouos univariate random variable is simply a density on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$ with respect to $\mathcal{L}^1$ (recall that $d\mathcal{L}^1(x) \equiv dx$).

Definition 51 (**Expected value**). *If $X \in Q(P)$ is a random variable, then the* **expected value of** *X, denoted $E(X)$, is defined as*

$$E(X) := \int_\Omega X dP.$$

Theorem 83 (**Some undergrad facts**). *If $X\colon \Omega \to \mathbb{R}$ is a random variable on $(\Omega, \mathcal{F}, P)$ then*

- *$\varphi(E(X)) \le E(\varphi(X))$ for any convex function $\varphi: \mathbb{R} \to \mathbb{R}$ such that $X \in L_1(P)$.*
- *$P(X \ge \alpha) \le \frac{E(X)}{\alpha}$ if $X \in \mathscr{N}$ and $\alpha \ge 0$.*
- *The cumulative distribution function, defined by $F(x) := P(X \le x)$, uniquely determines the distribution of X.*
- *$E(X) = \int_0^\infty P(X > t)dt = \int_0^\infty P(X \ge t)dt$ if $X \in \mathscr{N}$.*

If, in addition, the law of X has density f_X with respect to Lebsesque measure (i.e. $f_X = \frac{dPX^{-1}}{d\mathcal{L}^1}$), then

- *f_X is unique $\mathcal{L}^1$-a.e.*
- *$E(X) = \int_{\mathbb{R}} x f_X(x) dx$ if $X \in Q(P)$;*
- *$E(g(X)) = \int_{\mathbb{R}} g(x) f_X(x) dx$ if $g(X) \in Q(P)$ and g is measurable;*
- *If $T\colon \mathbb{R} \to \mathbb{R}$ is an invertible map from $\mathbb{R}$ onto $\mathbb{R}$ for which both T and T^{-1} are measurable and T^{-1} is continuously differentiable on $\mathbb{R}$, then the random variable $T(X)$ has a density, $f_{T(X)}$, with respect to Lebesque measure that satisfies*
$$f_{T(X)} = \big|(T^{-1})'\big| f_X \circ T^{-1}.$$

8.2 Application: likelihood of a inhomogeneous Poisson process

9 Integration to the limit

Section Assumption. *For the remainder of this section let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f_1, f_2, \ldots$ be measurable $\mathcal{F}/\mathcal{B}$ functions of Ω.*

Theorem 84 (**Fatou's Lemma).** *If $f_n \geq 0$ μ-a.e. then*

$$\int_\Omega \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \int_\Omega f_n d\mu$$

Theorem 85 (**Dominated Convergence Theorem).** *If $f_n \to f$ μ-a.e. and there exists a function $g \in L_1(\mu)$ such that $\sup_n |f_n| \leq g$ μ-a.e. then $f_n, f \in L_1(\mu)$ and*

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu.$$

Corollary 14 (**Bounded Convergence Theorem).** *If $\mu(\Omega) < \infty$, $f_n \to f$ μ-a.e. and there exists a constant $B < \infty$ such that $\sup_n |f_n| \leq B$. Then $f_n, f \in L_1(\mu)$ and*

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu.$$

Definition 52 (**Uniform Integrability).** *The sequence $f_1, f_2, \ldots$ is said to be* **uniformly integrable** *if*

$$\lim_{c\to\infty} \sup_n \int_\Omega |f_n| I_{\{|f_n|\geq c\}} d\mu = 0.$$

Theorem 86 (**Dilatation criterion for UI).** *If there exists an $\epsilon > 0$ such that $\sup_n \int_\Omega |f_n|^{1+\epsilon} d\mu < \infty$ then X_n are UI.*

Proof.

$$\begin{aligned}\int_\Omega |X_n| I_{\{|X_n|\geq c\}} d\mu &\leq \int_\Omega |X_n| \Big[\frac{|X_n|}{c}\Big]^\epsilon I_{\{|X_n|\geq c\}} d\mu \\ &\leq \frac{1}{c^\epsilon} \int_\Omega |X_n|^{1+\epsilon} d\mu.\end{aligned}$$

□

Theorem 87 (**UI theorem).** *If $\mu(\Omega) < \infty$, $f_n \to f$ μ-a.e. and the f_n are uniformly integrable, then $f_n, f \in L_1(\mu)$ and*

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu.$$

Theorem 88 (**UI converse).** *If*

1. $\mu(\Omega) < \infty$
2. $f_n \to f$ μ-a.e.
3. $f_n, f \in \mathscr{N} \cap L_1(\mu)$
4. $\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu$

then the f_n are uniformly integrable.

Theorem 89 (**Scheffé's theorem).** *Suppose P_n and P are probability measures on a measurable space $(\Omega, \mathcal{F})$ having densities δ_n and δ with respect to μ. If*

$$\delta_n \to \delta \ \mu\text{-a.e.}$$

then

$$\|P_n - P\|_{TV} := \sup_{A\in\mathcal{F}} |P_n(A) - P(A)| \leq \int_\Omega |\delta_n - \delta| d\mu \to 0.$$

Corollary 15. *If X is a random variable with a beta density $f_{\alpha,\beta}$ (with respect to Lebesque measure) given by*

$$f_{\alpha,\beta}(x) := \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x)$$

for $\alpha > 0$ and $\beta > 0$. Then the law of the random variable $(X - E(X))/sd(X)$ converges, in norm $\|\cdot\|_{TV}$, to a standard Gaussian distribution as $\alpha \to \infty$ and $\beta \to \infty$.

Theorem 90 (**Differentiability of $\int_\Omega f_t \, d\mu$).** *Let (a, b) be an open interval of $\mathbb{R}$ and $\{f_t\}_{t\in(a,b)}$ be a collection of functions on Ω. Suppose there exists $\Omega_0 \in \mathcal{F}$ such that:*

- $\mu(\Omega_0^c) = 0$;
- *For every $w \in \Omega_0$, $f_t(w)$ is differentiable at each $t \in (a, b)$;*
- *For every $w \in \Omega_0$, $\sup_{t\in(a,b)} \left|\frac{d}{dt} f_t(w)\right| \leq g(w)$;*
- $f_t \in L_1(\mu)$, $\forall t \in (a, b)$;
- $g \in L_1(\mu)$.

Then $\frac{d}{dt} f_t \in L_1(\mu)$, $\int_\Omega f_t \, d\mu$ is differentiable at each $t \in (a, b)$ and

$$\frac{d}{dt} \int_\Omega f_t \, d\mu = \int_\Omega \frac{d}{dt} f_t \, d\mu$$

at each $t \in (a, b)$.

Exercise 40. *Suppose that $f_1, f_2, \ldots$ and f are integrable and that $f_n \to f$ μ-a.e. Show that $\lim_n \int |f_n - f| d\mu = 0$ if and only if $\int |f_n| d\mu \to \int |f| d\mu$. Hint: for '$\Leftarrow$' study the proof of the DCT to show that $\limsup_n \int |f_n - f| d\mu \leq \int \limsup_n |f_n - f| d\mu$. In particular, show that $\int 2|f| d\mu - \int \limsup_n |f_n - f| d\mu \leq \int 2|f| d\mu - \limsup_n \int |f_n - f| d\mu$.*

Exercise 41 (**Sterling's formula for the Gamma function).** *The Gamma function is defined by the equality $\Gamma(r+1) := \int_0^\infty y^r e^{-y} dy$ for $r \in (0, \infty)$. Use the change of variable $z = (y - r)/\sqrt{r}$ to show that*

$$\rho_r := \frac{\Gamma(r+1)}{r^r e^{-r}\sqrt{r}} = \int_{-\sqrt{r}}^\infty e^{-\psi_r(z)} dz$$

where $\psi_r(z) := r\phi(z/\sqrt{r})$ with $\phi(u) := u - \log(1 + u)$. Next show that

$$\lim_{r\to\infty} \psi_r(z) = z^2/2 \text{ and } \psi_r(z) \geq c \min(z^2, \sqrt{r}|z|)$$

for some constant $c > 0$ (the largest admissible c is $\phi(1)$, but any c will work for the next step). Finally use the DCT to deduce that

$$\lim_{r\to\infty} \rho_r = \int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}.$$

Exercise 42 (**L^1 is complete).** *Let $f_1, f_2, \dots$ be integrable functions such that $\alpha_{m,n} := \int |f_n - f_m| d\mu$ tends to 0 as m and n tend to ∞. Show that there exists an integrable function f such that $\beta_n := \int |f - f_n| d\mu$ tends to 0 as n tends to ∞. Hint: inductively choose indices $n_k > n_{k-1}$ such that $\alpha_{m,n} \le 2^{-k}$ for all $m, n \ge n_k$ and set $f = \sum_{k=1}^{\infty} (f_{n_k} - f_{n_{k-1}})$ with $f_{n_0} = 0$.*

Exercise 43. *Suppose that $\Omega = (0,1)$, $\mathcal{F}$ is the Borel σ-field of Ω and μ is Lebesque measure on Ω. For $t \in T := (0,1)$, set $f_t(w) = I_{(0,t]}(w)$ and $J(t) := \int f_t d\mu$. Show that for each $t \in T$, $J(t)$ is differentiable at t but the derivative can not be computed under the integral sign, even though f'_t exists μ-a.e. and is integrable.*

9.1 Application: positive definite functions on Hilbert spaces

9.2 Application: complex generating function G_ν, characteristic function, moment generating function and Fourier transforms

Definition 53 (**Complex generating function).** *For any measure ν on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$ the function $G_\nu : \mathbb{C} \to \mathbb{C}$ defined by*

$$G_\nu(z) := \int_{\mathbb{R}} e^{zx}\, d\nu(x) \text{ for } z \in \mathbb{C}$$

is called the **complex generating function of** ν*. If X is a random variable then the complex generating function of X is defined as*

$$G_X(z) := E(e^{zX}).$$

Note that integration of functions taking values in $\mathbb{C}$ is simply done individually on the real and imaginary parts (treating $i = \sqrt{-1}$ as a constant). In particular, if Z is a complex random variable then $Z = X + iY$ where X, Y are both real random variables. Then, if X and Y are both in $Q(P)$ then we define

$$\int_\Omega Z dP := \int_\Omega X dP + i \int_\Omega Y dP.$$

Now, many of our results for integrating real random variable carry over to complex random variables.

Definition 54. *If X is a random variable then the* **moment generating function of** X *is defined as*

$$M_X(t) := G_X(t) \text{ for } t \in \mathbb{R}$$

and the **characteristic function of** X *is defined as*

$$\phi_X(t) := G_X(it) \text{ for } t \in \mathbb{R}.$$

Definition 55 (**Domain of G_X and M_X).** *If X is a random variable then* **the domain of** G_X *is defined as*

$$\mathfrak{G}_X := \{z \in \mathbb{C} \colon |G_X(z)| < \infty\}$$

and **the domain of** M_X *is defined as*

$$\mathfrak{M}_X := \{t \in \mathbb{R} \colon |M_X(t)| < \infty\}.$$

Theorem 91 (**Characterize $\mathfrak{M}_X$).** *If X is a random variable then $\mathfrak{M}_X$ is an interval containing 0 (perhaps empty, closed, open, half open or perhaps just the point 0).*

Theorem 92 (**$\mathfrak{G}_X$ is a cylinder above $\mathfrak{M}_X$).** *If X be a random variable then the domain of G_X is the cylinder above the domain of M_X. In particular*

$$\mathfrak{G}_X = \{z \in \mathbb{C} \colon \textit{real}(z) \in \mathfrak{M}_X\}.$$

Notice that the results on $\mathfrak{M}_X$ and and $\mathfrak{G}_x$ imply that M_X is only guaranteed to be finite at 0 whereas $\phi_\nu(t)$ is defined and finite for all $t \in \mathbb{R}$. In part, this explains the need to work with characteristic function rather than the moment generating function in that the latter is sometime degenerate.

Theorem 93 (**G_X is analytic on $\mathfrak{G}_X$).** *If X be a random variable then G_X is analytic on $\mathfrak{G}_X^o :=$ the interior of $\mathfrak{G}_X$.*

Theorem 94. *If X is a random variable then for any $z \in \mathfrak{G}_X^o$ one has that $X^n e^{zX} \in L_1(P)$ and*

$$G_X^{(n)}(z) = E(X^n e^{zX})$$

Theorem 95. *If X is a random variable then for any $t \in \mathfrak{M}_X^o$ one has that $X^n e^{tX} \in L_1(P)$ and*

$$M_X^{(n)}(t) = E(X^n e^{tX}).$$

Theorem 96. *If X is a random variable such that $M_X(t)$ has a right handed derivative at $t = 0$ then $X \in Q^+(P)$ and*

$$\frac{d^+ M_X(t)}{dt}\Big|_{t=0} = E(X).$$

10 Product measures and Fubini

Definition 56 (The section of a set). *For any set $A \in \Omega_1 \times \Omega_2$ the section of A determined by w_1 is defined as $A_{w_1} := \{w_2 \in \Omega_2 : (w_1, w_2) \in A\}$. Similarly, the section of A determined by w_2 is defined as $A_{w_2} := \{w_1 \in \Omega_2 : (w_1, w_2) \in A\}$.*

Definition 57 (The section of a function). *For any function $f : \Omega_1 \times \Omega_2 \to \Omega$ define the section of f determined by w_1 as $f(w_1, \cdot) : \Omega_2 \to \Omega$. Similarly, the section of f determined by w_1 is defined as $f(\cdot, w_2) : \Omega_1 \to \Omega$.*

Theorem 97 (Sections are measurable). *Let $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ be a measurable product space. Let A be a set in $\mathcal{F}_1 \otimes \mathcal{F}_2$ and f be a measurable $\mathcal{F}_1 \otimes \mathcal{F}_2/\mathcal{B}$ function. Then for any $w_1 \in \Omega$ and $w_2 \in \Omega_2$ then the sections $A_{w_1} \in \mathcal{F}_2$, $A_{w_2} \in \mathcal{F}_1$, $f(\cdot, w_2)$ is measurable $\mathcal{F}_1/\mathcal{B}$ and $f(w_1, \cdot)$ is measurable $\mathcal{F}_2/\mathcal{B}$.*

Theorem 98 (Product probablities). *Let P_1 and P_2 be probability measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. Then there exists a unique probability measure $P_1 \otimes P_2$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ such that*

$$P_1 \otimes P_2(A_1 \times A_2) = P_1(A_1)P_2(A_2)$$

for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

Theorem 99 (Fubinito). *Let P_1 and P_2 be probability measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. If $f : \Omega_1 \times \Omega_2 \to \bar{\mathbb{R}}$ is $\mathcal{F}_1 \otimes \mathcal{F}_2/\mathcal{B}$ measurable and $P_1 \otimes P_2$-quasi-integrable, then*

$$\int_{\Omega_1 \times \Omega_2} f dP_1 \otimes P_2 = \int_{\Omega_2} \Big[\int_{\Omega_1} f(\cdot, w_2) dP_1\Big] dP_2(w_2) \tag{44}$$

$$= \int_{\Omega_1} \Big[\int_{\Omega_2} f(w_1, \cdot) dP_2\Big] dP_1(w_1) \tag{45}$$

The inner integrals on the right hand side of (44) and (45) exist almost everywhere and are measurable, quasi-integrable functions of the sectioning variable.

Theorem 100 (Product measures). *Let μ_1 and μ_2 be σ-finite measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. Then there exists a unique σ-finite measure $\mu_1 \otimes \mu_2$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ such that*

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

for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

Theorem 101 (Fubini). *Theorem 99 (Fubinito) holds with the term "probability measure" replaced by "σ-finite measure".*

Corollary 16 (Useful re-wording of Fubini). *Suppose ν_1 and ν_2 are σ-finite measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. Let $f : \Omega_1 \times \Omega_2 \to \bar{\mathbb{R}}$ is $\mathcal{F}_1 \otimes \mathcal{F}_2/\mathcal{B}$ measurable. Consider the three integrals:*

$$D(f) := \int_{\Omega_1 \times \Omega_2} f d\nu_1 \otimes \nu_2$$

$$I_{1,2}(f) := \int_{\Omega_1} \Big[\int_{\Omega_2} f(w_1, \cdot) dP_2\Big] d\nu_1(w_1)$$

$$I_{2,1}(f) := \int_{\Omega_2} \Big[\int_{\Omega_1} f(\cdot, w_2) dP_1\Big] d\nu_2(w_2).$$

Then

$$D(f) \text{ is well defined} \Longrightarrow I_{1,2}(f) \text{ and } I_{2,1}(f) \text{ are well defined and } D(f) = I_{1,2}(f) = I_{2,1}(f).$$

Moreover

$$\begin{aligned} D(f) \text{ is well defined} &\Longleftrightarrow \text{either } D(f^+) \text{ or } D(f^-) \text{ is finite} \\ &\Longleftrightarrow \text{at least one of } I_{1,2}(f^+),\ I_{1,2}(f^-), \\ &\qquad I_{2,1}(f^+) \text{ or } I_{2,1}(f^-) \text{ is finite} \\ &\Longleftarrow \text{either } I_{1,2}(|f|) \text{ or } I_{2,1}(|f|) \text{ is finite.} \end{aligned}$$

Here is a nice corollary of Fubini

Corollary 17 (Integration term by term). *Suppose μ is a σ-finite measure*

1. *If $f_n \geq 0$ μ-a.e. then $\sum_{n=1}^{\infty} f_n \in Q(\mu)$ and*

$$\int_\Omega \sum_{n=1}^{\infty} f_n \, d\mu = \sum_{n=1}^{\infty} \int_\Omega f_n \, d\mu.$$

2. *If $\sum_{n=1}^{\infty} \int_\Omega |f_n| d\mu < \infty$ then $\sum_{n=1}^{\infty} |f_n| < \infty$ μ-a.e., $\sum_{n=1}^{\infty} f_n \in L_1(\mu)$ and*

$$\int_\Omega \sum_{n=1}^{\infty} f_n \, d\mu = \sum_{n=1}^{\infty} \int_\Omega f_n \, d\mu.$$

Corollary 18 (Using sectioning to compute product probabilities). *Suppose ν_1 and ν_2 are σ-finite measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. If $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$ then*

$$\nu_1 \otimes \nu_2(A) = \int_{\Omega_1} \nu_2(A_{w_1}) d\nu_1(w_1) = \int_{\Omega_2} \nu_1(A_{w_2}) d\nu_2(w_2).$$

Corollary 19 (Integrals that factor). *Suppose ν_1 and ν_2 are σ-finite measures on the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively. Let $f_1 : \Omega_1 \to \bar{\mathbb{R}}$ be $\mathcal{F}_1/\mathcal{B}$ measurable and $f_2 : \Omega_2 \to \bar{\mathbb{R}}$ be $\mathcal{F}_2/\mathcal{B}$. Then*

$$\int_{\Omega_1 \times \Omega_2} f_1(w_1) f_2(w_2) d\nu_1 \otimes \nu_2(w_1, w_2) = \prod_{i=1}^{2} \int_{\Omega_i} f_i(w_i) d\nu_i(w_i)$$

provided each f_i is nonnegative or each f_i is ν_i-integrable

Theorem 102 (Associativity of product measures). *If $(\Omega_i, \mathcal{F}_i, \nu_i)$ are σ-finite measure spaces for each $i = 1, 2, 3$ then $\nu_1 \otimes (\nu_2 \otimes \nu_3) = (\nu_1 \otimes \nu_2) \otimes \nu_3$ and $\mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes \mathcal{F}_3) = (\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes \mathcal{F}_3 = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \mathcal{F}_3$.*

The following theorem only covers the basics when working with finite dimensional product spaces. Once needs to get a bit more fancy in the definition when working with infinite product spaces

Definition 58 (Product measure of higher order). *Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be σ-finite measure spaces for each $i = 1, 2, \ldots, n$. The measure $\nu_1 \otimes \cdots \otimes \nu_n$ on $(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{F}_i)$, also denoted $\bigotimes_{i=1}^n \nu_i$, is defined as the σ-finite measure $\nu_1 \otimes (\nu_2 \otimes \nu_3)$ when $n = 3$ and extended recursively when $n > 3$.*

Theorem 103 (Higher order Fubini). *Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be σ-finite measure spaces for each $i = 1, 2, \ldots, n$. If $f : \prod_{i=1}^n \Omega_i \to \bar{\mathbb{R}}$ is $\bigotimes_{i=1}^n \mathcal{F}_i / \mathcal{B}$ measurable and $\nu_1 \otimes \cdots \otimes \nu_n$-quasi-integrable then*

$$\int_{\prod_i \Omega_i} f \, d\, \nu_1 \otimes \cdots \otimes \nu_n$$
$$= \int_{\Omega_{\pi_1}} \cdots \int_{\Omega_{\pi_n}} f(w_1, \ldots, w_n) d\nu_{\pi_n}(w_{\pi_n}) \cdots d\nu_{\pi_1}(w_{\pi_1})$$

for any permutation π of $\{1, 2, \ldots, n\}$ where the right hand side is interpreted as the iterated integral (starting with the inner most integral with respect to ν_{π_n}, then moving outward).

Corollary 20 (Borel σ-field and Lebesque measure). $(\mathbb{R}^d, \mathcal{B}^{\mathbb{R}^d}, \mathcal{L}^d) = (\mathbb{R}^d, \bigotimes_{i=1}^d \mathcal{B}^{\mathbb{R}}, \bigotimes_{i=1}^d \mathcal{L}^1)$

Corollary 21 (Integrate out the joint to get the marginal). *Let X_1, X_2 be two random variables on a probability space $(\Omega, \mathcal{F}, P)$. If the distribution of the random vector (X_1, X_2) has density $f_{X_1,X_2}(x_1, x_2)$ with respect to $\nu_1 \otimes \nu_2$ for two σ-finite measures ν_1, ν_2 on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$, then X_1 has density f_{X_1} with respect to ν_1 where*

$$f_{X_1}(x_1) := \int_{\mathbb{R}} f_{X_1,X_2}(x_1, x_2) d\nu_2(x_2).$$

10.1 Application: more random variable independence

Theorem 104 ('$\otimes$' means independence). *Let $X_1, \ldots, X_n$ denote independent random variables on some probability space $(\Omega, \mathcal{F}, P)$ with induced marginal distributions μ_i for X_i. Then*

$$P\big((X_1, \ldots, X_n) \in B\big) = \mu_1 \otimes \cdots \otimes \mu_n(B)$$

for all $B \in \mathcal{B}^{\mathbb{R}^n}$

Theorem 105 (Law of total probability for independent r.v.s). *If $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_k)$ are independent random vectors (i.e. $\sigma\langle X_1, \ldots, X_n\rangle$ is independent of $\sigma\langle Y_1, \ldots, Y_k\rangle$) defined on some probability space $(\Omega, \mathcal{F}, P)$ with induced distributions μ and ν on $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively. Then*

$$P\big[(X, Y) \in B\big] = \int_{\mathbb{R}^n} P\big[(x, Y) \in B\big] d\mu(x)$$

for all $B \in \mathcal{B}^{\mathbb{R}^{n+d}}$. Moreover

$$P\big[X \in A, (X, Y) \in B\big] = \int_A P\big[(x, Y) \in B\big] d\mu(x)$$

for all $A \in \mathcal{B}^{\mathbb{R}^n}$ and $B \in \mathcal{B}^{\mathbb{R}^{n+d}}$

Theorem 106 (Independence factors densities and expected values). *Let $X_1, \ldots, X_n$ be a sequence of independent random variables on some probability space $(\Omega, \mathcal{F}, P)$ with marginal densities f_i with respect to a σ-finite measure ν_i on $(\mathbb{R}, \mathcal{B}^{\mathbb{R}})$. Then the random vector $(X_1, \ldots, X_n)$ has density*

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

with respect to measure $\nu_1 \otimes \cdots \otimes \nu_n$. Moreover if $X_1, \ldots, X_n$ are all either non-negative or integrable then

$$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n).$$

Theorem 107 (Factoring densities implies independence). *Let $X_1, \ldots, X_n$ be a sequence of random variables on some probability space $(\Omega, \mathcal{F}, P)$. Suppose the distribution of the random vector $(X_1, \ldots, X_2)$ has density $f(x_1, \ldots, x_n)$ with respect to some product measure $\nu_1 \otimes \cdots \otimes \nu_n$ on $(\mathbb{R}^n, \mathcal{B}^{\mathbb{R}^n})$, where each ν_i is σ-finite. If*

$$f(x_1, \ldots, x_n) = g_1(x_1) \cdots g_n(x_n)$$

for non-negative functions g_i which are ⓜ $\mathcal{B}^{\mathbb{R}} / \mathcal{B}$ then $X_1, \ldots, X_n$ are independent.

10.2 Application: infinite sums of independent random variables

10.3 Application: Characterizing PX^{-1} with complex generating function G_X

The reason we need to wait till after Fubini to get these results is that we need the Law of total probability to get these results (Thm 105)

Theorem 108. *Suppose X is random variable. Then $G_X(it)$ as a function of $t \in \mathbb{R}$, characterizes the distribution of X.*

Theorem 109. *Suppose X is random variable and $\mathfrak{M}_X^o$ contains 0. Then $G_X(t)$ as a function of $t \in \mathbb{R}$, characterizes the distribution of X.*

Theorem 110 (Taylor series of $E(e^{tX})$). *Let $M_X(t)$ be the moment generating function for the random variable X. If there exists an $\epsilon > 0$ such that $(-\epsilon, \epsilon) \subset \mathfrak{M}_X$ then*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E(X^k), \quad \text{for } |t| < \epsilon.$$

Corollary 22 (Using the MGF to get moments).

- *If Z has a standard Gaussian distribution then $M_Z(t) = e^{t^2}$. Moreover, if k is even then $EZ^k = 1 \times 3 \times \cdots \times (k-1)$ and if k is odd then $EZ^k = 0$.*

- *If W has an exponential density $f_W(w) = \alpha e^{-\alpha w}$ then $M_W(t) = \frac{\alpha}{\alpha - t}$ whenever $t < \alpha$ and $EW^k = k!\alpha^{-k}$.*

- *If N is a Poisson random variable with density $f_N(r) = e^{-\lambda}\lambda^r/(r!)$ with respect to counting measure on $\{0, 1, 2, \ldots\}$, then $M_N(t) = e^{\lambda(e^t - 1)}$ and $E(N) = var(N) = \lambda$.*

Theorem 111 (When do moments characterize the distribution). *Let X and Y be random variables. If $E(X^k) = E(Y^k) =: \alpha_k$ for all $k \in \mathbb{N}$ and the radius of convergence of the power series $\sum_{k=1}^{\infty} \alpha_k u^k/k!$ is nonzero, then X has the same distribution as Y. use 304 notes 13-11*

It might be interesting here to include the log-normal example of the same moments but a different distribution.

Include an example which uses these results for proving Schoenberg and von Neumann's theorem for radially symmetric positive definite functions on an infinite dimensional Hilbert space (reference: Sterrneman and van Perlo-ten Kleij, *Spherical distributions: Schoenberg (1938) revisted*) .

10.4 Application: computing $E(X^a/Y^b)$ and $E(\log(X))$

Part III
Convergence of probability measures

11 Convergence almost everywhere

Section Assumption. *For the remainder of this section, unless stated otherwise, let $X, Y, X_1, X_2, \ldots,$ be random vectors taking values in $\mathbb{R}^d$ all defined on the same probability space $(\Omega, \mathcal{F}, P)$.*

Definition 59. *X_n converges to X almost everywhere (or almost surely or with probability one), written $X_n \xrightarrow{ae} X$, if $P(X_n \to X) = 1$.*

Notice that

$$\{X_n \not\to X\} = \bigcup_{\epsilon \in R} \big\{ \{|X_n - X| > \epsilon\} \text{ i.o.}_n \big\} \tag{46}$$

where $R := \{\epsilon \in \mathbb{R} : \ \epsilon > 0 \text{ and } \epsilon \text{ is rational}\}$. Therefore the sets $\{X_n \not\to X\}$ and $\{X_n \to X\}$ are both measurable whenever X_n, X are measurable. Moreover, equation (46) gives the following characterization of almost everywhere convergence.

Theorem 112 (i.o. characterization). *$X_n \xrightarrow{ae} X$ if and only if $P\big(\{|X_n - X| > \epsilon\} \text{ i.o.}_n\big) = 0$ for all $\epsilon > 0$.*

Proof. Let R be defined as in (46). Then

$$\begin{aligned}
&P(X_n \to X) = 1 \\
&\quad\iff P(X_n \not\to X) = 0 \\
&\quad\iff P\Big(\bigcup_{\epsilon \in R} \big\{ \{|X_n - X| > \epsilon\} \text{ i.o.}_n \big\} \Big) = 0 \\
&\quad\iff \forall \epsilon \in R,\ P\big(\{|X_n - X| > \epsilon\} \text{ i.o.}_n\big) = 0.
\end{aligned}$$

For all $\epsilon > 0$ consider an irrational $\tau > 0$ and let $\epsilon \in R$ satisfy $\epsilon < \tau$. Now

$$P\big(\{|X_n - X| > \tau\} \text{ i.o.}_n\big) \leq P\big(\{|X_n - X| > \epsilon\} \text{ i.o.}_n\big) = 0$$

□

Theorem 113 (Almost sure uniqueness of limits). *If $X_n \xrightarrow{ae} X$ and $X_n \xrightarrow{ae} Y$ then $X = Y$ almost everywhere.*

Proof. For any fixed $\omega \in \Omega$, if $X_n(\omega) \to X(\omega)$ and $X_n(\omega) \to Y(\omega)$ then $X(w) = Y(w)$. Therefore

$$\{X_n \to X\} \cap \{X_n \to Y\} \subset \{X = Y\}.$$

Since $P(X_n \to X) = P(X_n \to Y) = 1$ this implies $P(X = Y) = 1$. □

Theorem 114 (Cauchy criteria for convergence). *X_n converges a.e. to some real random vector if and only if for every $\epsilon > 0$*

$$\lim_n \lim_m P\Big(\sup_{n \leq p \leq m} |X_n - X_p| \geq \epsilon \Big) = 0. \tag{47}$$

Proof. For each n, m define

$$\begin{aligned}
I_{n,m} &:= \sup_{n \leq p \leq m} |X_n - X_p| \\
I_{n,\infty} &:= \sup_{n \leq p < \infty} |X_n - X_p| \\
II_n &:= \sup_{n \leq q, p < \infty} |X_q - X_p|.
\end{aligned}$$

Notice that (47) is equivalent to $\lim_n \lim_m P(I_{n,m} \geq \epsilon) = 0$. We need the following four facts.

Fact 1: $\lim_m I_{n,m} = I_{n,\infty}$.

Fact 2: $0 \leq I_{n,\infty} \leq II_n \leq 2 I_{n,\infty}$.

Fact 3: $\lim_n P(I_{n,\infty} > \epsilon) = P(\{I_{n,\infty} > \epsilon\} \text{ i.o.}_n)$.

Fact 4: $\lim_m P(I_{n,m} > \epsilon) = P(\lim_m \uparrow \{I_{n,m} > \epsilon\}) = P(I_{n,\infty} > \epsilon)$.

Fact 3 follows directly from Fatou's lemma and the fact that the monotonicity of $I_{n,\infty}$ implies $P\big(\{I_{n,\infty} > \epsilon\} \text{ i.o.}_n\big) = P\big(\{I_{n,\infty} > \epsilon\} \text{ a.a.}_n\big)$. Now we have

$$\begin{aligned}
&X_n \text{ converges a.e. to some real random vector} \\
&\quad\iff II_n \xrightarrow{ae} 0 \\
&\quad\iff I_{n,\infty} \xrightarrow{ae} 0, \qquad \text{by Fact 2} \\
&\quad\iff \forall \epsilon > 0,\ P\big(\{I_{n,\infty} > \epsilon\} \text{ i.o.}_n\big) = 0 \quad \text{by Theorem 112} \\
&\quad\iff \forall \epsilon > 0,\ \lim_n P(I_{n,\infty} > \epsilon) = 0 \quad \text{by Fact 3} \\
&\quad\iff \forall \epsilon > 0,\ \lim_n \lim_m P(I_{n,m} > \epsilon) = 0 \quad \text{by Fact 4} \\
&\quad\iff \forall \epsilon > 0,\ \lim_n \lim_m P(I_{n,m} \geq \epsilon) = 0.
\end{aligned}$$

□

Definition 60 (X-continuous functions). *Let $g \colon \mathbb{R}^d \to \mathbb{R}^k$ be a measurable function and define*

$$C_g := \{x \in \mathbb{R} \colon g \text{ is continuous at } x\};$$

C_g is called the continuity set of g. C_g is a Borel set (even if g is not measurable). Say that g is **X-continuous** *if*

$$P(X \in C_g) = 1.$$

Theorem 115 (Continuous mapping theorem). *Suppose $g \colon \mathbb{R} \to \mathbb{R}$ is measurable and X-continuous. Then*

$$X_n \xrightarrow{ae} X \Longrightarrow g(X_n) \xrightarrow{ae} g(X)$$

Proof. $\{X_n \to X\} \cap \{X \in C_g\} \subset \{g(X_n) \to g(X)\}$. □

Exercise 44. *Suppose $X_n \xrightarrow{ae} X$. Show that for every $\epsilon > 0$, $\lim_m P\big[\sup_{n \geq m} |X_n - X| > \epsilon\big] = 0$.*

Definition 61. *X_n is said to* **converge almost uniformly to** *X, written $X_n \xrightarrow{au} X$, if for every $\epsilon > 0$ there exists a measubale U_ϵ such that $P[U_\epsilon^c] \leq \epsilon$ and $X_n(\omega) \to X(\omega)$ uniformly for all $\omega \in U_\epsilon$.*

Exercise 45 (Egoroff's Theorem). *Show that $X_n \xrightarrow{ae} X$ if and only if $X_n \xrightarrow{au} X$. Hint: if $X_n \xrightarrow{ae} X$ then there exists a subsequence n_k such that $P(\sup_{n \geq n_k} |X_n - X| > 1/k) < 1/k^2$.*

11.1 Application: Kolmogorov's SLLN

As a warm up to Kolmogorov's strong law lets start with the assumption of finite second moments.

Theorem 116 (**SLLN when** $E(X^2) < \infty$)**.** *Let $X_1, X_2, \ldots$ be independent random variables, each distributed like some random variable X, all defined on the same probabilty space. Let $S_n := X_1 + \ldots + X_n$.*

- *If $E(X^2) < \infty$ then $S_n/n \xrightarrow{ae} E(X)$.*

The main technique here is to use Chebyshev's theorem and the first Borel-Cantelli lemma to get strong convergence of a subsequence, then analyze the discrepancies of the subsequences. This turns out to be useful for the full SLLN, but one needs to perform an extra trucation step.

Proof. Start by setting $\mu := E(X)$. Notice also that it is sufficient to only consider positive X. In particular if

$$\frac{S_{n,+}}{n} := \frac{X_1^+ + \cdots + X_n^+}{n} \xrightarrow{ae} E(X^+)$$
$$\frac{S_{n,-}}{n} := \frac{X_1^- + \cdots + X_n^-}{n} \xrightarrow{ae} E(X^-)$$

then $S_n/n = S_{n,+}/n - S_{n,-}/n \xrightarrow{ae} E(X^+) - E(X^-) = E(X)$ so the theorem follows. From now on assume X is positive.

By Chebyshev's theorem

$$P\Big[\big|S_n/n - E(S_n/n)\big| \geq \epsilon\Big] \leq \frac{\mathrm{var}(S_n/n)}{\epsilon^2} \leq \frac{E(X^2)}{\epsilon^2 n}.$$

If we consider a subsequence $n_k := \lceil \alpha^k \rceil$ where $\alpha \in (1, \infty)$ then $\sum_{k=1}^\infty \frac{E(X^2)}{\epsilon^2 n_k} < \infty$. By the first Borel-Cantelli lemma, for all $\epsilon > 0$

$$P\Big[\big|S_{n_k}/n_k - E(S_{n_k}/n_k)\big| \geq \epsilon \; i.o._k\Big] = 0$$

Therefore

$$S_{n_k}/n_k - \underbrace{E(S_{n_k}/n_k)}_{=\mu} \xrightarrow{ae} 0.$$

as $k \to \infty$. Now we use the positivity of X to show the full sequence S_n/n converges to μ. Notice that when $n_k \leq n \leq n_{k+1}$ we have that

$$\frac{S_{n_k}}{n_{k+1}} \leq \frac{S_n}{n} \leq \frac{S_{n_{k+1}}}{n_k} \tag{48}$$

so that

$$LHS = \frac{S_{n_k}}{n_{k+1}} = \frac{S_{n_k}}{n_k}\frac{n_k}{n_{k+1}} \xrightarrow{ae} \mu/\alpha$$
$$RHS = \frac{S_{n_{k+1}}}{n_k} = \frac{S_{n_{k+1}}}{n_{k+1}}\frac{n_{k+1}}{n_k} \xrightarrow{ae} \mu\alpha$$

where the above is true for every $\alpha \in (1,\infty)$, in particular for every $\alpha \in R := \{z : z \in \mathbb{Q}, z > 1\}$. Therefore

$$P\Big[\underbrace{\bigcap_{\alpha \in R}\big\{\mu/\alpha \leq \liminf_n S_n/n \leq \limsup_n S_n/n \leq \mu\alpha\big\}}_{=\{S_n/n \to \mu\}}\Big] = 1$$

□

Theorem 117 (**Kolmogorov's SLLN**)**.** *Let $X_1, X_2, \ldots$ be independent random variables, each distributed like some random variable X, all defined on the same probabilty space. Let $S_n := X_1 + \ldots + X_n$.*

- *If X is quasi-integrable then $S_n/n \xrightarrow{ae} E(X)$.*

Proof. The main idea is to mimic arguments for Theorem 116 but with an additional truncation argument. Again we can suppose without loss of generality that X is positive.

First consider the case $E(X) < \infty$. The idea is to analyze the truncated average T_n/n instead of S_n/n where

$$T_n := \sum_{i=1}^n X_i I_{\{X_i \leq i\}}.$$

Notice that for large i the terms $X_i I_{\{X_i \leq i\}}$ start to behave more like X_i. Moreover the small i terms in T_n/n are downweighted by $1/n$. Therefore once might expect T_n/n to behave like S_n/n for large n. To continue the proof we again we use Chebyshev

$$\begin{aligned} P\Big[\big|T_n/n - E(T_n/n)\big| \geq \epsilon\Big] &\leq \frac{\mathrm{var}(T_n/n)}{\epsilon^2} \\ &\leq \frac{1}{\epsilon^2 n^2}\sum_{i=1}^n E(X_i^2 I_{\{X_i \leq i\}}) \\ &\leq \frac{1}{\epsilon^2 n^2}\sum_{i=1}^n E(X_i^2 I_{\{X_i \leq n\}}) \\ &\leq \frac{E(X^2 I_{\{X \leq n\}})}{\epsilon^2 n}. \end{aligned} \tag{49}$$

We now notice that if we define the subsequence $n_k := \lceil \alpha^k \rceil$ where $\alpha \in (1, \infty)$ then the right hand side (above) is summable. In particular

$$\begin{aligned} \sum_{k=1}^\infty \frac{E(X^2 I_{\{X \leq n_k\}})}{n_k} &\underset{\text{Fubini}}{=} E\Big(X^2 \sum_{k=1}^\infty \frac{1}{n_k} I_{\{X \leq n_k\}}\Big) \\ &= E\Big(X^2\Big[0 + \cdots 0 + \frac{1}{n_j} + \frac{1}{n_{j+1}} + \cdots\Big]\Big) \end{aligned}$$

where j is the first index such that $X \leq n_j$, i.e. $\frac{X}{n_j} \leq 1$. Also notice the higher order terms can be bounded as follows

$$\frac{X}{n_{j+m}} = \frac{X}{\lceil \alpha^{j+m} \rceil} \leq \frac{X}{\alpha^{j+m}} = \frac{1}{\alpha^m}\frac{n_j}{\alpha^j}\frac{X}{n_j} \leq \frac{2}{\alpha^m}.$$

Therefore

$$X^2\Big[\frac{1}{n_j} + \frac{1}{n_{j+1}} + \cdots\Big] \leq X\Big[\frac{2}{\alpha^0} + \frac{2}{\alpha^1} + \cdots\Big] \tag{50}$$

Now since $E(X) < \infty$, the right hand side of (50) has finite expected value, and hence Borel-Cantelli gives

$$T_{n_k}/n_k - E(T_{n_k}/n_k) \xrightarrow{ae} 0 \tag{51}$$

as $k \to \infty$. Now if we can show that $E(T_{n_k}/n_k) = \mu + o(1)$ we can apply the same arguments as found in Theorem 116 to get

$$T_n/n \xrightarrow{ae} \mu \tag{52}$$

as $n \to \infty$.

Now we show $E(T_n/n) = \mu + o(1)$ and $T_n/n = S_n/n + o(1)$ with probability one. Notice that $E(T_n/n) = \frac{1}{n}\sum_{i=1}^n E(X_i I_{\{X_i \le i\}})$ where $\lim_i E(X_i I_{\{X_i\le i\}}) = \lim_i E(X I_{\{X\le i\}}) = E(X) = \mu$ by the DCT. Therefore Lemma 7 applies with $\mu_i := E(X_i I_{\{X_i\le i\}})$ to give

$$E(T_n/n) = \frac{1}{n}\sum_{i=1}^n \mu_i = \mu + o(1). \tag{53}$$

To finish lets analyze the terms in T_n versus the terms in S_n

$$P(X_i \ne X_i I_{\{X_i\le i\}}) = P(X_i > i).$$

Lemma 6 (below) gives that $\sum_{i=1}^\infty P(X_i > i) = E(\lceil X\rceil) < \infty$. Borel-Cantelli then gives $P(X_i \ne X_i I_{\{X_i\le i\}}\ i.o._i) = 0$ which implies that for the high-index terms in T_n are eventually exactly the same as in S_n. Therefore

$$T_n/n = S_n/n + o(1) \tag{54}$$

with probability one. Equations (51), (54) and (53) finish the proof of the case when $E(X) < \infty$.

Now consider the case $E(X) = \infty$. We simply show that $\liminf_n S_n/n = \infty$ with probability one (which allows us to conclude that $\liminf_n S_n/n = \limsup_n S_n/n = \lim_n S_n/n = \infty$ with probability one). Indeed

$$\begin{aligned}\liminf_{n\to\infty}\frac{S_n(w)}{n} &\ge \liminf_{n\to\infty}\frac{X_1(w)\wedge k + \cdots + X_n(w)\wedge k}{n}\\ &= E(X\wedge k), \quad \text{by the case above}\end{aligned}$$

for all $w \in A_k$ where $P(A_k) = 1$. Continuity from below in Big 3 implies $E(X\wedge k)\to\infty$. Therefore $\liminf_{n\to\infty} S_n(w)/n = \infty$ for all $w \in \cap_{k=1}^\infty A_k$ which has probability one. Therefore

$$S_n/n \xrightarrow{ae} \infty.$$

□

The following lemma was used in the above proof to analyze the difference between a truncated sum and the non-truncated sum.

Lemma 6 (**Expect the ceiling lemma**). *If X is a nonnegative random variable, then*

$$\sum_{i=0}^\infty P(X > i) = E(\lceil X\rceil). \tag{55}$$

Proof.

$$\sum_{i=0}^\infty P(X>i) = \sum_{i=0}^\infty E(I_{\{X>i\}}) \underset{\text{Fubini}}{=} E\Big(\underbrace{\sum_{i=0}^\infty I_{\{X>i\}}}_{=\lceil X\rceil}\Big).$$

□

The following lemma was used to show that the expected value of a truncated sum, in the most general proof of the SLLN, converges to the non-truncated expected value.

Lemma 7 (**Cesàr summation lemma**). *If $\mu_i \to \mu$ as $i\to\infty$, then $\big(\sum_{i=1}^n \mu_i\big)/n \to \mu$ as $n\to\infty$.*

Proof.

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^n \mu_i - \mu\Big| &\le \frac{1}{n}\sum_{i=1}^n |\mu_i-\mu|\\ &\le \frac{1}{n}\sum_{i=1}^m |\mu_i - \mu| + \sup_{i>m}|\mu_i-\mu|,\ m\le n\\ &=: I_{n,m} + II_m\end{aligned}$$

Taking a limit as $n\to\infty$ first one gets $\lim_n I_{n,m} = 0$, then take a limit as $m\to\infty$ to get $\lim_m II_m = \limsup_m |\mu_m - \mu| = 0$. □

11.2 Application: renewal theory

Theorem 118 (**Application to renewal theory**). *Let $X_1, X_2, \ldots$ be iid non-negative random variables with expected value $\mu \in (0,\infty]$. Let $S_n := X_1 + \ldots + X_n$ and for real numbers $t \ge 0$ set*

$$N_t := \sup\{n \ge 0\colon S_n \le t\}.$$

Then $N_t/t \to 1/\mu$ a.e..

Notice that N_t is the number of X_k's which fit between 0 and t. Since each X_k is expected to be μ, one might expect $\mu N_t \approx t$. Indeed, this is the heuristic behind the limit $N_t/t \to 1/\mu$ as $t\to\infty$.

Proof. First notice that $N_t < \infty$ for each real $t \ge 0$ but $N_t \xrightarrow{ae} \infty$ as $t\to\infty$. This follows since $S_n \uparrow \infty$ almost everywhere (because the SLLN gives $S_n/n \xrightarrow{ae} \mu$ and μ is assumed non-zero).

According to the definition of N_t we have $S_{N_t} \le t < S_{N_t+1}$ and therefore

$$\frac{S_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{S_{N_t+1}}{N_t+1}\frac{N_t+1}{N_t}. \tag{56}$$

Now letting $t\to\infty$ so that $N_t \xrightarrow{ae} \infty$ gives

$$\frac{t}{N_t} \xrightarrow{ae} \mu.$$

The result follows since $1/x$ is continuous on $x \in [0,\infty]$.

□

11.3 Application: Glivenko-Cantelli

11.4 Application: ergodic theory

12 Convergence in probability

Section Assumption. *For the remainder of this section, unless stated otherwise, let $X, Y, X_1, X_2, \ldots,$ be random vectors taking values in $\mathbb{R}^d$ all defined on the same probability space $(\Omega, \mathcal{F}, P)$.*

Definition 62. *X_n converges to X in probability, written $X_n \xrightarrow{P} X$, if*

$$\forall \epsilon > 0, \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

Theorem 119 (Almost sure uniqueness of limits). *If $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P} Y$ then $X = Y$ almost everywhere.*

In some sense, the difference between $\xrightarrow{ae}$ and $\xrightarrow{P}$ is given by the fact that Fatou's lemma is an inequality and not a strict identity. Recall one of the inequalities in Fatou's lemma: $\limsup P(A_n) \leq P(\limsup A_n)$. Since $\limsup_n A_n = \{A_n \ i.o._n\}$ we have

$$X_n \xrightarrow{ae} X \iff \text{for all } \epsilon,\ P(\limsup_n \{|X_n - X| \geq \epsilon\}) = 0$$

$$X_n \xrightarrow{P} X \iff \text{for all } \epsilon,\ \limsup_n P(\{|X_n - X| \geq \epsilon\}) = 0$$

This makes it clear that $\xrightarrow{ae}$ implies $\xrightarrow{P}$ (by Fatou) but that the otherway around is never possible

Theorem 120 (ae implies P).

$$X_n \xrightarrow{ae} X \text{ implies } X_n \xrightarrow{P} X. \tag{57}$$

There are cases where one can go backwards, but this either requires working with subsequences or additional assumptions such as monotonicity.

Theorem 121 (Monotonicity gives P implies $a.e.$). *If each X_n is a random variable and for almost every $w \in \Omega$, $X_n(w)$ is either nondecreasing in n or nonincreasing in n. Then*

$$X_n \xrightarrow{ae} X \iff X_n \xrightarrow{P} X.$$

Theorem 122 (Subsequences gives P implies $a.e.$). *$X_n \xrightarrow{P} X$ if and only if every subsequence $\{n_k\}_{k=1}^\infty$ contains a further subsequence $\{n_{k_\ell}\}_{\ell=1}^\infty$ such that $X_{n_{k_\ell}} \xrightarrow{ae} X$.*

As a corollary of the above theorem one gets that $X_n \xrightarrow{P} X$ implies there exists a subsequence $\{n_k\}_{k=1}^\infty$ such that $X_n \xrightarrow{ae} X$. One of the nice things about the subsequences theorem is that it allows us to generalize the theorems for taking a.e. limits under the integrals to the probability limits under the expectation. Here is an example of a theorem that we need later.

Theorem 123 (Probability Sandwich Theorem). *Suppose $0 \leq X_n \leq Y_n$ P-a.e., $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$, $E(Y_n) < \infty$ and $E(Y) < \infty$. If $E(Y_n) \to E(Y)$ then $E(X_n) < \infty$, $E(X) < \infty$ and*

$$E(X_n) \to E(X).$$

Proof. We start by showing the result under the stronger assumption that $X_n \xrightarrow{ae} X$, $Y_n \xrightarrow{ae} Y$. In this case Fatou gives

$$E(X) = E(\liminf_n X_n) \leq \liminf_n E(X_n). \tag{58}$$

Since the above $RHS \leq \liminf_n E(Y_n) = E(Y) < \infty$ equation (58) gives $E(X) < \infty$ (we also obviously have $E(X_n) < \infty$ by the inequality assumption). We also have that $0 \leq Y_n - X_n$ so again Fatou gives

$$\begin{aligned} E(Y) - E(X) &= E(Y - X) \\ &= E(\liminf_n (Y_n - X_n)) \\ &\leq \liminf_n E(Y_n - X_n) \\ &= E(Y) - \limsup_n E(X_n). \end{aligned}$$

Combined with equation (58) gives

$$E(X) \leq \liminf_n E(X_n) \leq \limsup_n E(X_n) \leq E(X). \tag{59}$$

Now we can use the subsequence theorem to weaken the assumption to $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. To show $E(X_n) \to E(X)$ proceed by contradiction and suppose there exists $\delta > 0$ and a subsequence n_k such that $|E(X_{n_k}) - E(X)| \geq \delta$. Now, by extracting a further subsequence n_{k_ℓ} where $X_{n_{k_\ell}} \xrightarrow{ae} X$ and $Y_{n_{k_\ell}} \xrightarrow{ae} Y$ we can conclude, by (59), that $E(X_{n_{k_\ell}}) \to E(X)$. This is a contradiction and therefore

$$E(X_n) \to E(X).$$

□

Theorem 124 (Cauchy criteria for $\xrightarrow{P}$). *X_n converges in P to some random variable if and only if*

$$\lim_n \lim_m \sup_{n \leq p \leq m} P\Big(|X_n - X_p| \geq \epsilon \Big) = 0. \tag{60}$$

if and only if

$$\lim_n \lim_m P\Big(|X_n - X_m| \geq \epsilon \Big) = 0. \tag{61}$$

Theorem 125 (Continuous mapping theorem for $\xrightarrow{P}$). *Suppose $g\colon \mathbb{R}^d \to \mathbb{R}^k$ is X-continuous. Then*

$$X_n \xrightarrow{P} X \Longrightarrow g(X_n) \xrightarrow{P} g(X)$$

Exercise 46 (Metrizing $\xrightarrow{P}$). *Let $\mathfrak{R}$ be the space of real-valued random variables on $(\Omega, \mathcal{F}, P)$. Let $d\colon \mathfrak{R} \times \mathfrak{R} \to [0,1]$ be defined by $d(X,Y) := E(|X - Y| \wedge 1)$. Show the following*

- *d is a pseudo metric on $\mathfrak{R}$*
- *$X_n \xrightarrow{P} X$ if and only if $d(X_n, X) \to 0$*
- *$\mathfrak{R}$ is speparable if $\mathcal{F}$ is countably generated (Hint: use Theorem 58 (part 1), Theorem 27 and Exercise 9 from the class notes).*
- *$\mathfrak{R}$ is complete.*

12.1 Stochastic order notation: O_p, o_p

We will come back to this later when we talk about convergence in distribution but it will be nice to get the notation set and a few restuls done.

Definition 63 (little o and big O). *Suppose $X_1, X_2, \dots$ are random vectors and $r_1, r_2, \dots$ are random variables all defined on $(\Omega, \mathcal{F}, P)$. Then*

$$X_n = o_p(r_n) \iff X_n/r_n \xrightarrow{P} 0$$
$$X_n = O_p(r_n) \iff \textit{the } \text{r.v.s } X_n/r_n \textit{ are tight}$$

Claim 1. *If $E|X_n|^p = O(1)$ for some $p > 0$ then $X_n = O_p(1)$.*

13 Convergence in L_p for $p \in [1, \infty)$

Warning. *We remind the reader that we are exclusively working with random variables or random vectors. In particular, the sample space Ω is assumed to have finite mass. Results for L_p convergence and L_p spaces with non-finite measures may be different.*

Section Assumption. *Thought this section assume $p \in [1, \infty)$, unless explicitly stated otherwise.*

Section Assumption. *Thought this section assume, unless stated otherwise, that $Y, X, X_1, X_2, \dots$ are all random vectors taking values in $\mathbb{R}^d$ defined on the same probability space $(\Omega, \mathcal{F}, P)$.*

Definition 64. $X_n \xrightarrow{L_p} X$ *if and only if* $E\big(|X_n - X|^p\big) \to 0$.

Theorem 126 (Almost sure uniqueness of limits). *If $X_n \xrightarrow{L_p} X$ and $X_n \xrightarrow{L_p} Y$ then $X = Y$ almost everywhere.*

Proof. First notice

$$\begin{aligned} |x+y|^p &= 2^p\Big|\frac{x+y}{2}\Big|^p \\ &\leq 2^p\Big(\frac{|x|^p + |y|^p}{2}\Big) \quad \text{by convexity of } |\cdot|^p \\ &\leq 2^{p-1}(|x|^p + |y|^p). \end{aligned} \tag{62}$$

Therefore

$$0 \leq E|X-Y|^p \leq 2^{p-1}\big(E|X-X_n|^p + E|Y-X_n|^p\big) \to 0. \tag{63}$$

□

Theorem 127 (Cauchy criteria for convergence). *X_n converges in L_p to some random variable if and only if*

$$\lim_n \lim_q E(|X_n - X_q|^p) = 0. \tag{64}$$

Proof. ($\Longrightarrow$) This follows immediately from the following version of (63)

$$0 \leq E|X_n - X_q|^p \leq 2^{p-1}\big(E|X_n - X|^p + E|X_q - X|^p\big).$$

($\Longleftarrow$) By Markov's inequality

$$P(|X_n - X_q| \geq \epsilon) \leq \frac{E|X_n - X_q|^p}{\epsilon^p}.$$

The Cauchy criterion for convergence in probability implies there exists a X such that $X_n \xrightarrow{P} X$. The subsequences theorem says that there exists a subsequence n_k such that $X_{n_k} \xrightarrow{ae} X$. Since $|X_n(\omega) - \cdot|^p$ is continuous for each $\omega \in \Omega$ we then have $|X_n - X_{n_k}|^p \xrightarrow{ae} |X_n - X|^p$ and therefore

$$\begin{aligned} E|X_n - X|^p &\leq \liminf_k E|X_n - X_{n_k}|^p \quad \text{by Fatou} \\ &\leq \limsup_k E|X_n - X_{n_k}|^p \\ &\leq \limsup_q E|X_n - X_q|^p. \end{aligned}$$

Taking $\lim_n$ on both sides gives the result.

□

Theorem 128 ($\xrightarrow{L_p}$ **implies** $\xrightarrow{P}$). *If $X_n \xrightarrow{L_p} X$ then $X_n \xrightarrow{P} X$*

Proof. This follows directly from Markov's theorem

$$P(|X_n - X| \geq \epsilon) \leq \frac{E|X_n - X|^p}{\epsilon^p} \to 0$$

as $n \to \infty$. □

13.1 L_p spaces of random vectors

Definition 65 (L_p **norm**). $\|X\|_p := \big[E(|X|^p)\big]^{1/p}$.

Definition 66 (L_p **space**). *Let L_p denote the collection of all random vectors $X : \Omega \to \mathbb{R}^d$ such that $\|X\|_p < \infty$.*

Theorem 129. *L_p is a linear space. In particular*

- *$X \in L_p$ and $c \in \mathbb{R} \Longrightarrow cX \in L_p$;*
- *$X, Y \in L_p \Longrightarrow X + Y \in L_p$.*

Proof. The first bullet follows trivially from the linear properties of expected value. For the second bullet, use inequality (62). □

Theorem 130 (Hölder). *Let X and Y be two random variables. If p, q are two positive numbers such that $\frac{1}{p} + \frac{1}{q} = 1$ then*

$$E(|X \cdot Y|) \leq \|X\|_p \|Y\|_q. \tag{65}$$

Proof. First recall our convection $0 \cdot \infty = 0$ for the right hand side of (65). Second notice that $\|X\|_p = 0$ implies $|X|^p = 0$ a.e. (by Theorem 72) which then implies $E|X \cdot Y| = 0$. Therefore inequality (65) holds if any one of the following is true: $\|X\|_p = 0$, $\|Y\|_q = 0$, $\|X\|_p = \infty$ or $\|Y\|_q = \infty$.

Now we may assume $\|X\|_p, \|Y\|_q \in (0, \infty)$. Define $Z := X/\|X\|_p$ and $W := Y/\|Y\|_q$. We need to show $E(|Z \cdot W|) \leq 1$. We first show Young's inequality

$$a^{w_1} b^{w_2} \leq w_1 a + w_2 b \tag{66}$$

for any $a, b \geq 0$ and $w_1, w_2 > 0$ such that $w_1 + w_2 = 1$. Young's inequality follows (after establishing the special cases when $a = 0$ or $b = 0$) by taking log of both sides and then using concavity to conclude that $w_1 \log(a) + w_2 \log(b) \leq \log(w_1 a + w_2 b)$. The inequality (66) now gives

$$\begin{aligned} E(|Z \cdot W|) &\leq E(|Z||W|), \text{ Cauchy-Schwarz for vectors} \\ &= E([|Z|^p]^{\frac{1}{p}}[|W|^q]^{\frac{1}{q}}) \\ &= E(\frac{1}{p}|Z|^p + \frac{1}{q}|W|^q). \end{aligned}$$

The result follows after noticing that $E(|Z|^p) = E(|W|^q) = 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. □

Notice that the above theorem holds for random variables which are not quasi-integrable. If it is known a priori that $\|X\|_p, \|Y\|_q < \infty$ then equation (65) can be extended to $|E(X \cdot Y)| \leq \|X\|_p \|Y\|_q$.

Theorem 131. *If $p \in (1,\infty)$ then $\|X\|_1 \le d\|X\|_p$.*

Proof. Use $Y = e_i$ in (65) where $e_1,\ldots,e_d$ is the standard basis for $\mathbb{R}^d$. □

Theorem 132. $q < p \Longrightarrow L_q \supset L_p$.

Proof. For $q < p$ Young's inequality (66) gives

$$|X|^q = \big[|X|^p\big]^{\frac{q}{p}} 1^{\frac{p-q}{p}} \le \frac{q}{p}|X|^p + \frac{p-q}{p}. \tag{67}$$

□

Theorem 133 ($\|\cdot\|_p$ **is a pseudo-norm).** *If $p \in [1,\infty)$ then $\|\cdot\|_p$ satisfies the following as a function over L_p:*

- $\|X\|_p \ge 0$
- $\|X\|_p = 0$ *implies* $X = 0$ *a.e.*
- $\|cX\|_p = |c|\|X\|_p$ *for any* $c \in \mathbb{R}$
- $\|X+Y\|_p \le \|X\|_p + \|X\|_p$ *(Minkowski's inequality).*

Proof. It has already been noticed that $\|X\|_p = 0$ implies $|X|^p = 0$ a.e. (by Theorem 72) which then implies $X = 0$ a.e.. The third bullet is trivial from the linear properties of E. To prove Minkowski notice

$$\begin{aligned} E(|X+Y|^p) &= E\big(|X+Y||X+Y|^{p-1}\big) \\ &\le E\big(|X||X+Y|^{p-1}\big) + E\big(|Y||X+Y|^{p-1}\big) \\ &= E\big(|X||X+Y|^{q(p-1)/q}\big) + E\big(|Y||X+Y|^{q(p-1)/q}\big) \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Notice $p = q(p-1)$ which implies $|X+Y|^{q(p-1)/q} = |X+Y|^{p/q}$. Therefore

$$\begin{aligned} E(|X+Y|^p) &\le E\big(|X||X+Y|^{p/q}\big) + E\big(|Y||X+Y|^{p/q}\big) \\ &\le \|X\|_p \||X+Y|^{p/q}\|_q + \|Y\|_p \||X+Y|^{p/q}\|_q \\ &\le \big(\|X\|_p + \|Y\|_p\big)\||X+Y|^{p/q}\|_q \end{aligned}$$

The proof now follows after noticing that $E(|X+Y|^p)/\||X+Y|^{p/q}\|_q = \|X+Y\|_p$. □

Definition 67 (distances in L_p). *The distance between $X \in L_p$ and $X \in L_p$ is defined as $d_p(X,Y) := \|X-Y\|_p$*

Theorem 134 (d_p is a pseudo-metric on L_p). *If $p \in [1,\infty)$ then for all $X,Y,Z \in L_p$,*

- $d_p(X,Y) \ge 0$
- $d_p(X,Y) = 0$ *implies* $X = Y$ *a.e.*
- $d_p(X,Y) = d_p(Y,X)$
- $d_p(X,Y) \le d_p(X,Z) + d_p(Z,Y)$.

Proof. This follows directly from Theorem 133. □

Theorem 135 (Continuity of $\|\cdot\|_p$). *For any $X,Y \in L_p$, $|\|X\|_p - \|Y\|_p| \le d_p(X,Y)$.*

Proof. By Minkowski $\|X\|_p = \|X-Y+Y\|_p \le \|X-Y\|_p + \|Y\|_p$. Therefore

$$\|X\|_p - \|Y\|_p \le \|X-Y\|_p.$$

By symmetry we also have $\|Y\|_p - \|X\|_p \le \|X-Y\|_p$ which is sufficient to finish the proof. □

Theorem 136 (L_p is closed). *If $\|X_n - X\|_p \to 0$ and $X_n \in L_p$ then $X \in L_p$.*

Proof. This follows since (62) implies

$$|X|^p \le 2^{p-1}|X_n|^p + 2^{p-1}|X_n - X|^p$$

where the expected value of the right hand side is finite (for large enough n). □

Theorem 137 (L_p is complete). *Cauchy sequences (in the d_p metric) converge to a member in L_p*

Proof. This was already proved in Theorem 127. □

Theorem 138 (L_p is separable). *If the σ-field $\mathcal{F}$ is countably generated then there exists a countable dense subset of L_p.*

Proof. See Exercise 47. □

Exercise 47. *(a) Suppose $\mathcal{F}_0$ is a field generating $\mathcal{F}$. Show that the set of $\mathcal{F}_0$-simple functions are dense in L_p. (b) Show that L_p is separable (i.e. there exists a countable) if $\mathcal{F}$ is countably generated. Hint: use Exercise 9 and Theorem 28.*

13.2 L_p convergence theorem

Recall the definition of uniformly integrable.

Definition 68 (UI). *If X_n is a sequence of random vectors such that*

$$\lim_{c\to\infty} \sup_n E\big(|X_n| I_{\{|X_n| \ge c\}}\big) = 0$$

then $X_1, X_2, \ldots$ are said to be **uniformly integrable** *(UI).*

Theorem 139 (L_p convergence theorem). *Suppose $X_n \in L_p$ for all n. The following are equivalent*

1. $X_n \xrightarrow{L_p} X$
2. $X_n \xrightarrow{P} X$ *and* $E|X_n|^p \to E|X|^p < \infty$
3. $X_n \xrightarrow{P} X$ *and* $|X_n|^p$*'s are UI*

Proof. (*1.* $\Longrightarrow$ 2.) By Theorem 128 we know that $X_n \xrightarrow{P} X$. We also know that $E|X|^p < \infty$ since L_p is closed by Theorem 136. Finally by the continuity result in Theorem 135 we know that $\|X_n\|_p \to \|X\|_p < \infty$.

(*2.* $\Longrightarrow$ 1.) We use the Sandwich theorem. Define

$$Y_n := 2^{p-1}(|X_n|^p + |X|^p), \qquad Y := 2^p|X|^p.$$

Notice the following facts

- $Y_n \xrightarrow{P} Y$ by the Continuous mapping theorem;
- $Y_n, Y \in L_1$ since $X_n, X \in L_p$;
- $0 \le |X_n - X|^p \le Y_n$ by equation (62);
- $E(Y_n) \xrightarrow{P} E(Y)$ by assumption;
- $|X_n - X|^p \xrightarrow{P} 0$ by the Continuous mapping theorem.

Therefore Sandwich Theorem 123 applies and gives $E|X_n - X|^p \to 0$.

(*2.* $\Longrightarrow$ 3.) We use the old UI Converse Theorem 88 modified for convergence in probability. Proceeding by contradiction suppose the $|X_n|^p$'s are *not* UI. In particular,

$$\lim_{c\to\infty} \sup_n E\big(|X_n|^p I_{\{|X_n|^p \ge c\}}\big) \neq 0.$$

Now there exists a $\delta > 0$ and a a sequence of real numbers $c_k \to \infty$ such that

$$\sup_n E\big(|X_n|^p I_{\{|X_n|^p \ge c_k\}}\big) > \delta.$$

for all k. For each k one can now choose n_k such that

$$E\big(|X_{n_k}|^p I_{\{|X_{n_k}|^p \ge c_k\}}\big) > \delta. \tag{68}$$

However, there exists a further subsequence $|X_{n_{k_\ell}}|^p$ such that $|X_{n_{k_\ell}}|^p \xrightarrow{ae} |X|^p$ and $E|X_{n_{k_\ell}}|^p \xrightarrow{ae} E|X|^p < \infty$. Applying the old UI Converse Theorem 88 we then get the $|X_{n_{k_\ell}}|^p$'s are UI so that

$$\lim_{c\to\infty} \sup_\ell E\big(|X_{n_{k_\ell}}|^p I_{\{|X_{n_{k_\ell}}|^p \ge c\}}\big) = 0$$

which contradicts equation (68). Therefore the $|X_n|^p$'s are UI.

(*3.* $\Longrightarrow$ 2.) This follows by our old UI Theorem 87. In particular, by taking subsequences and applying Theorem 87 we have $E|X|^p < \infty$. To show $E|X_n|^p \to E|X|^p$ we proceed by contradiction and suppose there exists a subsequence n_k and a $\delta > 0$ such that

$$\big|E|X_{n_k}|^p - E|X|^p\big| \ge \delta. \tag{69}$$

By taking subsequences we get that $|X_{n_{k_\ell}}|^p \xrightarrow{ae} |X|^p$ where the $|X_{n_{k_\ell}}|^p$'s are UI. Now again by Theorem 87 we have that $E|X_{n_{k_\ell}}|^p \to E|X|^p < \infty$ which contradicts (69). □

13.3 Special geometry of L_2

Notice that for any two $X, Y \in L_2$ one can use Hölder's inquality to get

$$E(|X \cdot Y|) \le \|X\|_2 \|Y\|_2. \tag{70}$$

In particular $E(X \cdot Y)$ is defined and finite for any two $X, Y \in L_2$. This motivates the following definition of an inner product in L_2

Definition 69. *For any $X, Y \in L_2$ the inner product of X and Y is defined as*

$$\langle X, Y\rangle := E(X \cdot Y) \tag{71}$$

Theorem 140 (Properties of $\langle\cdot,\cdot\rangle$). *For all $X, Y, Z \in L_2$*

1. $\langle X, X\rangle \ge 0$
2. $\langle X, X\rangle > 0$, *except when $X = 0$ a.e.*
3. $\langle X, Y\rangle = \langle Y, X\rangle$
4. $\langle X, Y + \alpha Z\rangle = \langle X, Y\rangle + \alpha\langle X, Z\rangle$ *when $\alpha \in \mathbb{R}$*
5. *If $X_n \xrightarrow{L_2} X$ then $\langle X_n, Y\rangle \to \langle X, Y\rangle$ for all Y.*

Proof. The first four statement follow trivially from properties of expected value. For the last that

$$|\langle X_n, Y\rangle - \langle X, Y\rangle| = |\langle X_n - X, Y\rangle| \le \|X_n - X\|_2 \|Y\|_2 \to 0.$$

□

Theorem 141 (Pythagorean). *For any $X, Y \in L_2$,*

$$\|X + Y\|_2^2 = \|X\|_2^2 + 2\langle X, Y\rangle + \|Y\|_2^2.$$

Proof. This follows trivially from properties of expected value but it's useful to notice that this follows directly from Theorem 140 and the fact that $\langle X, X\rangle = \|X\|_2^2$:

$$\begin{aligned}
\|X + Y\|_2^2 &= \langle X + Y, X + Y\rangle \\
&= \langle X + Y, X\rangle + \langle X + Y, Y\rangle, \text{ by linearity} \\
&= \langle X, X + Y\rangle + \langle Y, X + Y\rangle, \text{ by symmetry} \\
&= \langle X, X\rangle + \langle X, Y\rangle + \langle Y, X\rangle + \langle Y, Y\rangle, \text{ by linearity} \\
&= \|X\|_2^2 + 2\langle X, Y\rangle + \|Y\|_2^2.
\end{aligned}$$

□

Theorem 142 (Parallelogram). *For any $X, Y \in L_2$,*

$$\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2\|X\|_2^2 + 2\|Y\|_2^2.$$

Proof. Add the following two equations:

$$\begin{aligned}
\|X + Y\|_2^2 &= \|X\|_2^2 + 2\langle X, Y\rangle + \|Y\|_2^2 \\
\|X - Y\|_2^2 &= \|X\|_2^2 - 2\langle X, Y\rangle + \|Y\|_2^2.
\end{aligned}$$

□

Definition 70 (Orthogonal). *$X \in L_2$ is said to be orthogonal to $Y \in L_2$, denoted $X \perp Y$, if $\langle X, Y\rangle = 0$.*

Theorem 143 (Projection theorem). *Let S be a closed linear subspace of L_2 and let $Y \in L_2$. Then there exists an almost surely unique member of S, denoted $\mathcal{P}_S Y$, such that*

$$\|Y - \mathcal{P}_S Y\|_2 = \inf\{\|Y - X\|_2 \colon X \in S\}. \tag{72}$$

Moreover, $\mathcal{P}_S Y$ is characterized by the property that $\mathcal{P}_S Y \in S$ and

$$X \perp (Y - \mathcal{P}_S Y) \text{ for all } X \in S. \tag{73}$$

Proof. (*Existance*) Choose $X_n \in S$ such that $\|Y - X_n\|_2 \to \inf\{\|Y - X\|_2 \colon X \in S\}$. The sequence X_n is Cauchy. In particular, by the Parallelogram equality

$$\|X_n - Y + X_m - Y\|_2^2 + \|X_n - X_m\|_2^2 = 2\|X_n - Y\|_2^2 + 2\|X_m - Y\|_2^2.$$

If we set $X = \frac{1}{2}(X_n + X_m)$ then

$$4\|X - Y\|_2^2 + \|X_n - X_m\|_2^2 = 2\|X_n - Y\|_2^2 + 2\|X_m - Y\|_2^2.$$

Since $X \in S$ one has $\|X - Y\|_2^2 \geq \inf^2$. Therefore

$$\|X_n - X_m\|_2^2 \leq \underbrace{2\|X_n - Y\|_2^2 + 2\|X_m - Y\|_2^2 - 4\inf^2}_{\to 0 \text{ as } n,m\to\infty}.$$

Indeed, X_n is a Cauchy sequence. Therefore there exists a limit, call it $\mathcal{P}_S Y$, such that $X_n \xrightarrow{L_2} \mathcal{P}_S Y$ which must be in S since it's closed. Let check that $\|\mathcal{P}_S Y - Y\|_2 = \inf$. Indeed, $\inf \leq \|\mathcal{P}_S Y - Y\|_2 \leq \|\mathcal{P}_S Y - X_n\|_2 + \|X_n - Y\|_2 \to \inf$.

(*Uniqueness*) To show uniqueness use the same trick. Suppose $X \in S$ and $\|X - Y\|_2 = \inf$. Then by the same Parallelogram equality

$$\|X - Y + P_S Y - Y\|_2^2 + \|X - \mathcal{P}_S Y\|_2^2 = 2\|X - Y\|_2^2 + 2\|\mathcal{P}_S Y - Y\|_2^2.$$

Since $W := \frac{1}{2}(X + P_S Y) \in S$ we then have that

$$\begin{aligned}\|X - \mathcal{P}_S Y\|_2^2 &= 2\inf^2 + 2\inf^2 - 4\|W - Y\|_2^2 \\ &\leq 2\inf^2 + 2\inf^2 - 4\inf^2 = 0.\end{aligned}$$

Therefore $\|X - \mathcal{P}_S Y\|_2^2 = 0$ and hence $\mathcal{P}_S Y$ is almost surely unique.

((73) $\Rightarrow$ (72)) If $\mathcal{P}_S Y$ satisfies (73) then for every $X \in S$ one has

$$\|X - Y\|_2^2 = \|\mathcal{P}_S Y - Y\|_2^2 + \|\underbrace{X}_{\in S} - \underbrace{\mathcal{P}_S Y}_{\in S}\|_2^2$$

Therefore to minimize the left hand side choose $X = \mathcal{P}_S Y$ and get equation (72).

((72) $\Rightarrow$ (73)) Let $X \in S$ such that $X \neq 0$ a.e. (otherwise (73) is trivially true). The idea is to define

$$f(c) := \|Y - (\mathcal{P}_S Y - cX)\|_2^2$$

and notice that the minimum of $f(c)$, call it $c_{\min}$, can be computed in two ways. The first way is to notice that $f(c)$ is minimized at $c_{\min} = 0$ by equation (72). Therefore $c_{\min} = 0$. The other way to compute $c_{\min}$ is to use the fact that

$$f(c) = \|Y - \mathcal{P}_S Y\|_2^2 + 2c\langle Y - \mathcal{P}_S Y, X\rangle + c^2\|X\|_2^2$$

which is minimized at $c_{\min} = \langle Y - \mathcal{P}_S Y, X\rangle / \|X\|_2^2$ (note we are using the fact that $X \neq 0$ a.e). The two ways to compute $c_{\min}$ must be the same, or else $\mathcal{P}_S Y$ would not be unique. Setting two ways to compute $c_{\min}$ equal to each other gives

$$0 = \frac{\langle Y - \mathcal{P}_S Y, X\rangle}{\|X\|_2^2}$$

which proves (73). ☐

Definition 71 (**Orthonormal set**)**.** *A set of random vectors $\{X_i : i \in I\} \subset L_2$ are said to be orthonormal if $\langle X_i, X_j\rangle = 0$ for all $i \neq j$ and $\|X_i\|_2 = 1$ for all i.*

Definition 72 (**Infinite sums in L_2**)**.** *Suppose $Y, X_1, X_2, \ldots$ are members of L_2 and c_i are real numbers. We write $Y = \sum_{i=1}^\infty c_i X_i$ as short hand for $\|Y - \sum_{i=1}^N c_i X_i\|_2 \to 0$ as $N \to \infty$.*

Theorem 144 (**Computing a projection**)**.** *Let $X_1, X_2, \ldots$ denote a countable orthonormal set of random variables. Let S denote the collection of L_2 limits of finite linear combinations of the X_i's. Then S is a closed linear subset of L_2 and for any $Y \in L_2$ the projection of Y onto S is computed as follows*

$$\mathcal{P}_S Y = \sum_{i=1}^\infty \langle X_i, Y\rangle X_i.$$

Proof. (*S is closed and linear*) Lets first see why S is linear. Let $W, Z \in S$. There must exist W_n and Z_n which are finite linear combinations of the X_i's such that $W_n \xrightarrow{L_2} W$ and $Z_n \xrightarrow{L_2} Z$. Immediately one now has $aW + bZ \in S$ since $aW_n + bZ_n \xrightarrow{L_2} aW + bZ$ by Minkowskis inequality. To see that S is closed suppose $Z_n \in S$ converges to some Z. Let Z_n' be a finite linear combination of the X_i's such that $\|Z_n - Z_n'\|_2 \leq 1/n$. Then

$$\|Z_n' - Z\|_2 \leq \|Z_n - Z\|_2 + 1/n.$$

The right hand side converges to zero and therefore $Z_n' \xrightarrow{L_2} Z$, which establishes that $Z \in S$.

(*$\sum_i \langle X_i, Y\rangle X_i$ exists in L_2*) Exercise 48 shows that the projection of Y down to the span of $X_1, \ldots, X_n$ is computed as $\sum_{i=1}^n \langle X_i, Y\rangle X_i$. Moreover this projection decreases length so that

$$\textstyle\sum_{i=1}^n \langle X_i, Y\rangle^2 = \left\|\sum_{i=1}^n \langle X_i, Y\rangle X_i\right\|_2^2 \leq \|Y\|_2^2 < \infty.$$

This holds for each n which implies $\sum_{i=1}^\infty \langle X_i, Y\rangle^2 < \infty$. This allows us to conclude that $\sum_{i=1}^n \langle X_i, Y\rangle X_i$ is a Cauchy sequence. In particular,

$$\lim_n \lim_q \left\|\textstyle\sum_{i=n+1}^q \langle X_i, Y\rangle X_i\right\|_2^2 \leq \lim_n \textstyle\sum_{i=n+1}^\infty \langle X_i, Y\rangle^2 \to 0.$$

Therefore $\sum_i \langle X_i, Y\rangle X_i$ exists in L_2.

($\mathcal{P}_S Y = \sum_{i=1}^\infty \langle X_i, Y\rangle X_i$) We use characterization (73). Since $\sum_{i=1}^n \langle X_i, Y\rangle X_i \xrightarrow{L_2} \sum_{i=1}^\infty \langle X_i, Y\rangle X_i$ we have

$$\underbrace{\left\langle X_k, Y - \textstyle\sum_{i=1}^n \langle X_i, Y\rangle X_i\right\rangle}_{=0 \text{ for all } n} \to \left\langle X_k, Y - \textstyle\sum_{i=1}^\infty \langle X_i, Y\rangle X_i\right\rangle$$

by item 5 in Theorem 140. Therefore $Y - \sum_{i=1}^\infty \langle X_i, Y\rangle X_i$ is orthogonal to all X_k and hence to all S (again using 5 in Theorem 140). Therefore $\mathcal{P}_S Y = \sum_{i=1}^\infty \langle X_i, Y\rangle X_i$. ☐

Definition 73 (**Orthonormal basis**)**.** *A set of random variables $\{X_i : i \in I\} \subset L_2$ are said to be an orthonormal basis if finite linear combinations of X_i are dense in L_2.*

Theorem 145 (Properties of an ONB). *If $\{X_i : i \in I\} \subset L_2$ is an orthonormal set then the following are equivalent:*

1. $\{X_i : i \in I\} \subset L_2$ *is a basis*
2. $Y = \sum_{i=1}^{\infty} \langle X_i, Y\rangle X_i$ *for all* $Y \in L_2$
3. $\langle Y, Z\rangle = \sum_{i=1}^{\infty} a_i b_i$ *where* $a_i := \langle X_i, Y\rangle$ *and* $b_i := \langle X_i, Z\rangle$ *for all* $Y, Z \in L_2$
4. $\|Y\|_2^2 = \sum_{i=1}^{\infty} a_i^2$ *where* $a_i := \langle X_i, Y\rangle^2$ *for all* $Y \in L_2$.

Proof. (1. $\iff$ 2.) The direction $\Leftarrow$ is obvious. For $\Rightarrow$ it follows immediately from Theorem 144. Indeed, if the X_i's form a basis then the set of L_2 limits of finite linear combinations of the X_i's, S, simply equals L_2. Therefore

$$Y = \mathcal{P}_{L_2} Y = \sum_{i=1}^{\infty} \langle X_i, Y\rangle X_i.$$

(2. $\Longrightarrow$ 3.) Let $a_i := \langle X_i, Y\rangle$ and $b_i := \langle X_i, Z\rangle$. Then

$$\begin{aligned}\langle Y, Z\rangle &= \big\langle \textstyle\sum_{i=1}^{\infty} a_i X_i, \sum_{j=1}^{\infty} b_j X_j\big\rangle \\ &= \lim_n \big\langle \textstyle\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{\infty} b_j X_j\big\rangle, \text{ by Theorem 140} \\ &= \lim_n \textstyle\sum_{i=1}^{n} a_i b_i.\end{aligned}$$

(3. $\Longrightarrow$ 4.) Trivial.

(4. $\Longrightarrow$ 2.) Let S be the L_2 limits of the finite linear combinations of the X_i's. Considering Theorem 144 it will be sufficient to show that $\|\mathcal{P}_S Y - Y\|_2^2 = 0$. By property (73) of projections we have

$$\|Y\|_2^2 = \|Y - \mathcal{P}_S Y + \mathcal{P}_S Y\|_2^2 = \|Y - \mathcal{P}_S Y\|_2^2 + \|\mathcal{P}_S Y\|_2^2.$$

But now

$$\begin{aligned}\|Y - \mathcal{P}_S Y\|_2^2 &= \|Y\|_2^2 - \|\mathcal{P}_S Y\|_2^2 \\ &= \textstyle\sum_{i=1}^{\infty} \langle X_i, Y\rangle^2 - \|\sum_{i=1}^{\infty} \langle X_i, Y\rangle X_i\|_2^2 \\ &= 0.\end{aligned}$$

□

Theorem 146 (When does L_2 have an ONB). *If the σ-field $\mathcal{F}$ is countably generated then there exists an orthonormal basis of L_2.*

Proof. Theorem 138 says there exists a dense countable subset. Let $Y \in L_2$ and suppose $Y_n \xrightarrow{L_2} Y$ where each Y_n is in the dense countable subset. Let $\{X_i : i \in \mathbb{N}\}$ be a gram-schmidt orthogonalization of the dense countable subset. Then for each Y_n there exists a finite linear combination which of the X_i's which equal Y_n and therefore the all the finite linear combinations of the X_i's are dense. This is the definition of an ONB. □

Theorem 147 (L_2 is a Hilbert space). *By identifying every element in L_2 with the equivalence class of a.e.-modifications of random variables, the space L_2 with inner product defined as in (71) is a Hilbert space. In particular, L_2 is a complete linear vector space with strictly positive inner product. If, in addition, the σ-field $\mathcal{F}$ is countably generated then L_2 is a separable Hilbert space.*

Exercise 48. *For this exercise you are not allowed to use Theorem 144 or any of the results after.*

1. *Let S be a closed linear subset of L_2. Show that projection decreases length: $\|\mathcal{P}_S Y\|_2^2 \le \|Y\|_2^2$ for all $Y \in L_2$.*
2. *Let $X_1, \ldots, X_n$ be an finite set of orthonormal random variables. Let S_n denote the set of linear combinations of the X_i's. Show that $\mathcal{P}_{S_n} Y = \sum_{i=1}^{n} \langle X_i, Y\rangle X_i$.*

Exercise 49. *Show that if Gaussian random variables converge to a random variable with probability one, then that random variable is also Gaussian and the convergence also holds in L_2.*

13.4 Application: Gaussian conditional expected value as a projection

Ignoring, for the time being, that we technically have not defined conditional expectation yet, we can use projections to analyze finite dimensional Gaussian conditional expectation. Indeed, Gaussian conditional expectation is simply projection within L_2 on to the closed linear space of linear combinations of the observations. In this section we will informally (not rigorously) see the consequences of this and relate our theorems for deriving this conditional.

Suppose $(\Omega, \mathcal{F}, P)$ is rich enough to support $n+1$ random variables $Y, X_1, \ldots, X_n$ which are jointly Gaussian with $E(X_k) = E(Y) = 0$. In particular suppose the density the random vector $(Y, X_1, \ldots, X_n)$ on $\mathbb{R}^{n+1}$ is proportional to $\exp(-(y,x)^t \Sigma^{-1}(y,x)/2)$ where Σ is a positive definite matrix and $x \in \mathbb{R}^n$. Since Gaussian random variables have finite variance it is clear that each $X_i \in L_2$. By examining the undergraduate characterization of the conditional distribution of Y given $X_1, \ldots, X_n$ as a ratio densities, it becomes clear (after completing the square) that the conditional expectation of $E(Y|X_1, \ldots, X_n)$ must of the form $c_1 X_1 + \cdots + c_n X_n$. Let S denote the closed linear subspace in L_2 of finite linear combinations of the X_i's.

Now we can see why $E(Y|X_1, \ldots, X_n) = \mathcal{P}_S Y$. Let to save notational space we simply write $X := (X_1, \ldots, X_n)$. From our undergraduate understanding of conditional expected value we have that for any $W \in S$

$$\begin{aligned}E\big[Y - W\big]^2 &= E\big[Y - E(Y|X) + E(Y|X) - W\big]^2 \\ &= E\big[Y - E(Y|X)\big]^2 + E\big[E(Y|X) - W\big]^2 \end{aligned} \tag{74}$$

where we have used some undergrad facts like:

$$\begin{aligned}&E\Big\{\big[Y - E(Y|X)\big]\big[E(Y|X) - W\big]\Big\} \\ &\quad= E_X E_{Y|X}\Big\{\big[Y - E(Y|X)\big]\big[E(Y|X) - W\big]\Big\} \\ &\quad= E_X\Big\{\big[E(Y|X) - W\big]\underbrace{E_{Y|X}\big[Y - E(Y|X)\big]}_{=0}\Big\}\end{aligned}$$

which requires $W \in S$. Anyway, the upshot of (74) is that

$$E(Y|X) = \arg\min_{W \in \mathcal{S}} E\big[Y - W\big]^2$$

so that $E(Y|X) = \mathcal{P}_S Y$. Notice that (73) now shows that

$$Y - \mathcal{P}_S Y \perp X_1, \ldots, X_n$$

which, in turn, implies

$$\operatorname{var}(Y - \mathcal{P}_S Y) = \operatorname{var}(Y - \mathcal{P}_S Y|X) = \operatorname{var}(Y|X). \tag{75}$$

The last equality follows since $\mathcal{P}_S Y$ is a linear combination of the X_i's and is therefore a constant conditional on $X_1, \ldots, X_n$.

Notice a few nice consequences. First suppose I want to simulate from $Y|X$ but I only have an algorithm that can do two things: simulate a new pair (Y^*, X^*) with the same law as (Y, X) and compute the projection $E(Y|X)$ for any Y, X. To simulate $Y|X$ notice that $Y|X \sim \mathcal{N}\big(E(Y|X), \operatorname{var}(Y|X)\big)$. Therefore all I need is to simulate $Z \sim \mathcal{N}(0, 1)$ and then $E(Y|X) + Z \operatorname{std}(Y|X)$ will suffice as a conditional simulation from $Y|X$. But notice that (75) tells us that $Y^* - \mathcal{P}_{S^*} Y^*$ will have the same variance (and expected value) as $Z \operatorname{std}(Y|X)$ where (Y^*, X^*) is an independent simulation of the data and response. Therefore

$$E(Y|X) + Y^* - \mathcal{P}_{S^*} Y^*$$

serves as a conditional simulation of $Y|X$ when we only needed to be able to simulate from the joint measure (Y, X) and compute $E(Y|X)$.

Lets also use the fact that we know how to compute projections to easily compute $E(Y|X)$. We need to set down a orthonormal basis of S. This is easily done by $Z = \Sigma_{xx}^{-1/2} X$. Indeed $\langle Z_i, Z_j \rangle = \delta_{ij}$. Now projection is easy by (144)

$$E(Y|X) = \mathcal{P}_S Y = \sum_{i=1}^n \langle Z_i, Y \rangle Z_i = \Sigma_{yx} \Sigma_{xx}^{-1} X$$

where the last line is from Exercise 50. It is also easy to compute the conditional variance

$$\begin{aligned}
\operatorname{var}(Y|X) &= \operatorname{var}(Y - \mathcal{P}_S Y) \\
&= \|Y\|_2^2 - 2\sum_{i=1}^n \langle Z_i, Y \rangle^2 + \sum_{i=1}^n \langle Z_i, Y \rangle^2 \\
&= \|Y\|_2^2 - \sum_{i=1}^n \langle Z_i, Y \rangle^2 \\
&= \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}.
\end{aligned}$$

Exercise 50. *Using the notation above*

1. *Show the vector of coefficients* $\langle Z_1, Y \rangle \ldots \langle Z_n, Y \rangle$ *equals* $\Sigma_{xx}^{-1/2} \Sigma_{xy}$.
2. *Show that* $\sum_{i=1}^n \langle Z_i, Y \rangle Z_i = \Sigma_{yx} \Sigma_{xx}^{-1} X$.
3. *Show that* $\sum_{i=1}^n \langle Z_i, Y \rangle^2 = \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}$

13.5 Application: Hilbert spaces associated with a random field

14 Weak convergence in L_p when $p \in (1, \infty)$

In this class we will only prove one theorem from this section: Riesz's theorem for L_2. One of the reasons is that many of the other proofs use the Radon-Nikodym theorem which we have not proved yet. Moreover, we will not have occasion to use this theory much. Indeed, the main reason for this section is to give context to Riesz's theorem (understanding that it naturally lives in the larger theory of weak converges) and that weak convergence in L_p is, I think, the right way to understand the next section on convergence in distribution.

Definition 74 (**Dual space of L_p**). *The dual space L_p is the set of all continuous linear functionals on L_p.*

Definition 75 (**Weak convergence in L_p**). *X_n converges weakly to X in L_p, denoted $X_n \rightharpoonup_p X$, if $X_n, X \in L_p$ and if $f(X_n) \to f(X)$ for all functions f in the dual space of L_p. We use the shorthand $X_n \rightharpoonup X$ to denote $X_n \rightharpoonup_2 X$.*

One of the crucial differences between convergence in L_p and weak convergence in L_p is that $X_n \xrightarrow{L_2} X$ implies $\|X_n\|_p \to \|X\|_p$ whereas $X_n \rightharpoonup_p X$ does not. The following theorem shows that the two notions of convergence are indeed equivalent when $\|X_n\|_p \to \|X\|_p$.

Theorem 148 (**Relate with $\xrightarrow{L_p}$**). *Suppose $p \in (1, \infty)$ and $X, X_1, X_2, \ldots \in L_p$. Then $X_n \xrightarrow{L_p} X$ if and only if $X_n \rightharpoonup_p X$ and $\|X_n\|_p \to \|X\|_p$.*

Theorem 149 (**Riesz: dual of L_p is L_q**). *Suppose $p \in (1, \infty)$. Let $f : L_p \to \mathbb{R}$ be continuous and linear. Then there exists an almost surely unique $Y \in L_q$, where $\frac{1}{q} + \frac{1}{p} = 1$ such that $f(X) = \langle X, Y\rangle$ for all $Y \in L_q$.*

Theorem 150 (**Almost sure uniqueness of limits**).

Theorem 151 (**Cauchy criterion for $\rightharpoonup_p$**). *If for all $Y \in L_p$ the sequence of numbers $\{\langle X_n, Y\rangle\}_n$ is Cauchy then there exists an almost surely unique $X \in L_p$ such that and $X_n \rightharpoonup_p X$.*

Theorem 152 (**Weak compactness for L_p**). *If $X_n \in L_p$ is a bounded, i.e. there exists a C such that $\|X_n\|_p \le C$ for all n, then there exists a weakly convergent subsequence in L_p.*

14.1 Special case of L_2

Theorem 153 (**Riesz for L_2**). *Let $f : L_2 \to \mathbb{R}$ be continuous and linear. Then there exists an almost surely unique $Y \in L_2$ such that $f(X) = \langle Y, X\rangle$ for all $X \in L_2$*

Proof. Start by defining N to be the null space

$$N := \{X \in L_2 \colon f(X) = 0\}.$$

The idea is that the orthogonal complement of N is one dimensional and is spanned by a single $Y \in L_2$. Scaling Y appropriately will give $f(X) = \langle Y, X\rangle$ for all $X \in L_2$.

Choose a non-zero $Y \notin N$ (if there is no such Y then $f(X) = 0$ for all $X \in L_2$ and the theorem is trivially true). Additionally assume Y is orthogonal to N by projecting it out, if necessary. If $f(X) = \langle Y, X\rangle$ we at least need $f(Y) = \|Y\|_2^2$, so scale Y to satisfy this. Now we can apply the following decomposition for all $X \in L_2$

$$X = cY + Z \tag{76}$$

where $c := f(X)/\|Y\|_2^2$ and $Z := X - cY$. This will imply $Z \in N$ which will then imply cY is the projection of X down to the space spanned by Y. This will be sufficient to establish the proof since it will imply $\frac{\langle Y,X\rangle}{\|Y\|_2^2} = c = \frac{f(X)}{\|Y\|_2^2}$.

To see why $Z \in N$ notice

$$f(Z) = f(X - cY) = f(X) - \tfrac{f(X)}{\|Y\|_2^2} f(Y) = 0$$

where the last line uses the fact that $f(Y)/\|Y\|_2^2 = 1$. Now, since we picked Y to be orthogonal to N we have $Y \perp Z = X - cY$. This implies that cY is the projection of X into the space generated by Y. This establishes that $\frac{\langle Y,X\rangle}{\|Y\|_2^2} = c = \frac{f(X)}{\|Y\|_2^2}$. Therefore there exist an $Y \in L_2$ such that $f(X) = \langle Y, X\rangle$. To establish uniqueness let $\widetilde{Y} \in L_2$ which also satisfies $f(X) = \langle \widetilde{Y}, X\rangle$. Therefore $\langle \widetilde{Y} - Y, X\rangle = 0$ for all $X \in L_2$. To finish simply choose $X := \widetilde{Y} - Y$ to get $\|\widetilde{Y} - Y\|_2^2 = 0$.

□

14.2 Application: L_1 Wasserstein CLT with Stein's method

15 Convergence in Distribution

This is basically weak convergence in L_1 of the densities of X_n.

Definition 76 (Definition of weak convergence of probability measures). *Let P_n (for $n \in \mathbb{N}$) and P be probability measures on $(S, \mathcal{B}^S)$ for some metric space S. Then P_n converges weakly to P as $n \to \infty$, written $P_n \rightsquigarrow P$, if $\int_S f dP_n \to \int_S f dP$ for every bounded and continuous real function $f : S \to \mathbb{R}$.*

Definition 77 (Convergence in distribution for random vectors). *If X_n and X are random vectors in $\mathbb{R}^d$ then $X_n \rightsquigarrow X$ if and only if $Ef(X_n) \to Ef(X)$ as $n \to \infty$ for all bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$.*

Notice that notion of convergence does not require the random vectors X_n and X are all defined on the same probability space.

Theorem 154 (Portmanteau I). *Let S be a metric space and let $P, P_1, P_2, \ldots$ be probability measures on $(S, \mathcal{B}^S)$. Then the following statements are equivalent*

1. $P_n \rightsquigarrow P$
2. $\limsup_n P_n(F) \leq P(F)$ *for all closed* $F \subset S$.
3. $\liminf_n P_n(G) \geq P(G)$ *for all open* $G \subset S$.
4. $\lim_n P_n(A) = P(A)$ *for all* $A \in \mathcal{B}^S$ *such that* $P(\partial A) = 0$.

Theorem 155 (Portmanteau II). *Let X, X_1, X_2 be random variables (where X_n is defined on $(\Omega_n, \mathcal{F}_n, P_n)$ and X is defined on $(\Omega, \mathcal{F}, P)$). Let $F_n(x) := P_n(X_n \leq x)$ and $F(x) := P(X \leq x)$ and F_n^{-1}, F^{-1} be the corresponding left-continuous inverse cdf as defined in (34). Then the following are equivalent.*

1. $X_n \rightsquigarrow X$
2. $F_n(x) \to F(x)$, $\forall x$ *s.t.* $P(X = x) = 0$.
3. $F_n^{-1}(u) \to F^{-1}(u)$, $\forall u$ *s.t.* $F^{-1}(u)$ *is continuous at* u.

Theorem 156 (Distributional uniqueness of limits). *Let S be a metric space and let $Q, P, P_1, P_2, \ldots$ be probability measures on $(S, \mathcal{B}^S)$. If $P_n \rightsquigarrow P$ and $P_n \rightsquigarrow Q$ then $Q = P$ on $(S, \mathcal{B}^S)$.*

Theorem 157 (Subsequence criterion). *Let S be a metric space and let $P, P_1, P_2, \ldots$ be probability measures on $(S, \mathcal{B}^S)$. If for every subsequence n_k there exists a further subsequence n_{k_j} such that $P_{n_{k_j}} \rightsquigarrow P$ as $j \to \infty$, then $P_n \rightsquigarrow P$ as $n \to \infty$.*

Theorem 158 (P implies $\rightsquigarrow$). *Suppose X, X_n, Y_n are random vectors all defined on the same probability space $(\Omega, \mathcal{F}, P)$. If $Y_n \rightsquigarrow X$ and $|X_n - Y_n| \xrightarrow{P} 0$ then $X_n \rightsquigarrow X$.*

Theorem 159 (Skorokhod gives $\rightsquigarrow$ implies *a.e.*). *If $X_n \rightsquigarrow X$ then there exists on some probability space random variables $X_1^*, X_2^*, \ldots$ and X^* such that $X_n^* \sim X_n$ for each n, $X^* \sim X$ and $X_n^* \xrightarrow{ae} X^*$*

Skorokhod is extremely useful for weakening results for $\xrightarrow{ae}$ to $\rightsquigarrow$. The following theorem is a classic example.

Theorem 160 (Continuous mapping theorem). *Suppose X, X_n are random variables and $g\colon \mathbb{R} \to \mathbb{R}$ is X-continuous. Then*

$$X_n \rightsquigarrow X \Longrightarrow g(X_n) \rightsquigarrow g(X).$$

Here are a few examples where we can extend the results for passing limits under expected values for convergence in distribution. This technique almost universally applies, just so long as the conditions and conclusions (besides $X_n \rightsquigarrow X$) are in terms of marginal distributional properties of X_n and X.

Theorem 161 (Fatou). *Suppose X, X_n are random variables such that $X_n \geq 0$ a.e. and $X_n \rightsquigarrow X$. Then*

$$E(X) \leq \liminf_n E(X_n).$$

Theorem 162 (UI). *Suppose X, X_n are random variables such that the X_n's are UI and $X_n \rightsquigarrow X$. Then $X_n, X \in L_1$ and*

$$E(X_n) \to E(X) \quad \text{and} \quad E|X_n| \to E|X|.$$

Theorem 163 (Δ-method). *Let $X_1, X_2, \ldots$ and Z be random variables such that*

$$c_n(X_n - x_0) \rightsquigarrow Z$$

as $n \to \infty$ where $x_0 \in \mathbb{R}$ and c_n is a sequence of positive numbers tending to ∞. If $g : \mathbb{R} \to \mathbb{R}$ is differentiable at x_0 then

$$c_n(g(X_n) - g(x_0)) \rightsquigarrow g'(x_0)Z.$$

Theorem 164 (Portmanteau III). *Let X_n and X be random vectors. Then the following statements are equivalent*

1. $X_n \rightsquigarrow X$
2. $Ef(X_n) \to Ef(X)$ *for all bounded X-continuous* f.
3. $Ef(X_n) \to Ef(X)$ *for all bounded Lipschitz* f.

Theorem 165 (Portmanteau IV). *Let X_n and X be random vectors. Then the following statements are equivalent*

1. $X_n \rightsquigarrow X$
2. $Ef(X_n) \to Ef(X)$ *for all* $f \in \{\sin(x \cdot t)\colon t \in \mathbb{R}^d\}$.
3. $Ef(X_n) \to Ef(X)$ *for all* $f \in \{\cos(x \cdot t)\colon t \in \mathbb{R}^d\}$.
4. $Ef(X_n) \to Ef(X)$ *for all* $f \in \{e^{ix \cdot t}\colon t \in \mathbb{R}^d\}$.

Theorem 166 (Prohorov's theorem, aka conditions for sequenctial compactness). *Let X_n and X be a random vectors in $\mathbb{R}^k$. Then*

1. *If $X_n \rightsquigarrow X$ then $X_n = O_p(1)$;*
2. *If $X_n = O_p(1)$ then there exists a subsequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$ for some X.*

15.1 Application: CLT with characteristics functions

15.2 Application: edgeworth expansions

16 Mixing convergence types

More on stochastic order notation: O_p, o_p.

Theorem 167 (**Slutsky's theorem**). *Suppose X_n and Y_n are real random variables such that $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{P} c$ where c is a finite constant. Then*

$$(X_n, Y_n) \rightsquigarrow (X, c) \text{ in } \mathbb{R}^2. \tag{77}$$

In paritcular, by applying the definition of weak convergence one gets

1. $X_n + Y_n \rightsquigarrow X + c$
2. $X_n Y_n \rightsquigarrow cX$
3. $X_n Y_n \rightsquigarrow X/c$ *provided* $c \neq 0$.

Proof. To show (77) use $\mathcal{F}_X = \{\text{Lipschitz continuous funs }\}$ in theorem ??. To show the consequences use $\mathcal{F}_X = \{\text{bdd continuous funs }\}$ □

Exercise 51. *needs editing Show the following statements:*

1. *If X_n and X are random vectors in $\mathbb{R}^k$ then $X_n \xrightarrow{P} c$ for a constant c if and only if $X_n \rightsquigarrow c$;*
2. *If $X_n \rightsquigarrow X$ and $d(X_n, Y_n) \xrightarrow{P} 0$ then $X_n \rightsquigarrow Y$*
3. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $(X_n, Y_n) \xrightarrow{P} (X, Y)$.*

Part IV
Conditional probability

This part of the notes develops conditional expectation and probability. To motivate all this deep mathematics notice the following paradox. Suppose X and Y are independent random variables, each with a uniform distribution on $(0,1)$. For a Borel set B consider how to compute $P[X \in B | X = Y]$. There are a number of ways one might approach this problem. First, define $Z := X - Y$, use transformation of densities to find the density $f_{X,Z}(x,z)$, and then set $P[X \in B | X = Y] = P[X \in B | Z = 0] = \frac{f_{X,Z}(x,0)}{f_Z(0)}$ where f_Z is the marginal density of Z. Another way is to define $W := X/Y$ and then set $P[X \in B | X = Y] = P[X \in B | W = 1] = \frac{f_{X,W}(x,1)}{f_W(1)}$. The big problem is that these give different answers. Even more astounding is that, in some sense, neither approach is fundamentally wrong. Add a detailed example where one observes a random field with either multiplicative numerical truncation and one with additive numerical truncation

Here is a broad overview of what we are doing. Our basic strategy will be to define conditional expected value, $E^{\mathcal{B}}X$, with respect to a sub σ-field $\mathcal{B} \subset \mathcal{F}$. Once we have conditional expected value then we can get conditional probability by the expected value of indicators. The main tool for $E^{\mathcal{B}}X$ is the Radon-Nikodym derivative. In particular define suppose $X \in \mathscr{N}$ and let ν be defined by

$$\nu[B] = \int_B X dP$$

for all $B \in \mathcal{B}$. Now we use the Radon-Nikodym theorem to get the existance of $d\nu/dP$, which is $\mathcal{B}$ measurable, such that

$$\int_B X dP = \int_B \frac{d\nu}{dP} dP.$$

$d\nu/dP$ then serves as $E^{\mathcal{B}}X$.

17 Radon-Nikodym derivatives

Definition 78 (Absolute continuity and singularity). *Let ν and μ be measures on $(\Omega, \mathcal{A})$.*

- *ν and μ are said to be* **singular**, *denoted $\nu \perp \mu$, if and only if there exists a set $A \in \mathcal{A}$ such that*

$$\nu(A^c) = 0 = \mu(A).$$

- *ν is said to be* **absolutely continuous with respect to** *μ, denoted $\nu \ll \mu$, if and only if*

$$\nu(A) = 0 \text{ for every } \mathcal{A}\text{-set } A \text{ for which } \mu(A) = 0.$$

The following theorem and proof is probably one of my favorite in all of probability/measure theory.

Theorem 168 (Radon-Nikodym). *Let μ and ν be two σ-finite measures on $(\Omega, \mathcal{A})$. If $\nu \ll \mu$ then there exists a measurable function $\frac{d\nu}{d\mu} \in \mathscr{N}$ such that*

$$\nu[A] = \int_A \frac{d\nu}{d\mu} d\mu$$

for all $A \in \mathcal{A}$. Moreover, $\frac{d\nu}{d\mu}$ is μ-unique.

Proof. Notice first that μ-uniqueness follows directly from the uniqueness of densities Theorem 12.

The existent of $\frac{d\nu}{d\mu}$ is trivially true if either μ or ν is identically zero. So, from now on, suppose both are not identically zero. We are in search of $\frac{d\nu}{d\mu}$. The non-trivial assumption and the σ-finite assumption on both μ and ν allows us to apply our world view Theorem 82 and establish the existence of two probability measures P and Q on $(\Omega, \mathcal{A})$ such that $\frac{d\mu}{dP}$ and $\frac{d\nu}{dQ}$ both exist and map into $(0,\infty)$. Moreover, Exercise 37 says $\frac{dP}{d\mu} = 1/\frac{d\mu}{dP}$ and $\frac{dQ}{d\nu} = 1/\frac{d\nu}{dQ}$. Then notice that the chain rule Theorem 78 says that if $\frac{dQ}{dP}$ exists then so does $\frac{d\nu}{d\mu}$ and can be computed as follows

$$\frac{d\nu}{d\mu} = \frac{d\nu}{dQ}\frac{dQ}{dP}\frac{dP}{d\mu}.$$

Therefore we have reduced the problem to finding $\frac{dQ}{dP}$ where Q and P are probability measures such that $Q \ll P$ (this last condition follows since the existence of $\frac{d\mu}{dP}$ and $\frac{dQ}{d\nu}$ implies $Q \ll \nu \ll \mu \ll P$). Consider the probability measure $W = (Q+P)/2$. The construction of W ensures $Q \ll W$ and $P \ll W$. The idea is that we'll use Reisz to get $\frac{dP}{dW}$ and $\frac{dQ}{dW}$ then show that $\frac{dQ}{dP} = \frac{dQ}{dW}/\frac{dP}{dW}$.

Define the following functionals over over $L^2(W) := \{X\colon \int X^2 dW < \infty\}$

$$f_P(X) := \int X dP \quad \text{and} \quad f_Q(X) := \int X dQ.$$

By Exercise 52, both f_P and f_Q are continuous linear functionals over $L^2(W)$. By Riesz's Theorem 153 there exists random variables p and q such that

$$f_P(X) = \langle p, X\rangle = \int pX dW$$
$$f_Q(X) = \langle q, X\rangle = \int qX dW$$

for all $X \in L^2(W)$. In the case when $X = I_A$ we have

$$P[A] = f_P(I_A) = \int_A p\, dW$$
$$Q[A] = f_Q(I_A) = \int_A q\, dW.$$

Therefore $p = \frac{dP}{dW}$ and $q = \frac{dQ}{dW}$. In what follows I will analyze the ratio q/p but I want to avoid ∞/∞. I can ensure this is avoided by noticing that since P, Q and W are all probability measures, p and q must takes values in $[0,\infty)$ with W-probability one.

Therefore I may, and do, modify p and q over W-null sets so that they never take on the value ∞.

To finish we show that modifying q/p on the appropriate set serves as $\frac{dQ}{dP}$. Define $N := \{p = 0\}$ and set

$$\frac{dQ}{dP} := \begin{cases} q/p & \text{on } N^c \\ 0 & \text{on } N. \end{cases} \tag{78}$$

Now $\frac{dQ}{dP}$ has the right properties since

$$\begin{aligned}
\int_A \frac{dQ}{dP} dP &= \int_{A\cap N^c} (q/p)\, dP \\
&= \int_{A\cap N^c} (q/p)p\, dW, \text{ by Theorem 77} \\
&= \int_{A\cap N^c} q\, dW, \text{ since } (q/p)p = q \text{ on } N^c. \\
&= Q[A \cap N^c] + Q[A \cap N], \\
&\qquad \text{since } Q[A \cap N] = 0 \text{ by } P[N] = 0 \text{ and } Q \ll P \\
&= Q[A].
\end{aligned}$$

□

Definition 79. *For two measures ν, μ on $(\Omega, \mathcal{A})$ define the notation $\nu \lll \mu$ to mean that $\nu \ll \mu$ and μ is σ-finite.*

Theorem 169 (Radon-Nikodym*). *Theorem 168 is still true under the weaker assumption $\nu \lll \mu$.*

Proof. The problem with this case is that we are no longer guaranteed that Q exists (in the proof of Theorem 168). But we still have the existence of P and $\frac{dP}{d\mu}$. Moreover for this P we have $\nu \ll \mu \ll P$. Therefore all we need is to show that there exists $\frac{d\nu}{dP}$ under the assumption $\nu \ll P$ and we can then construct $\frac{d\nu}{d\mu}$ by

$$\frac{d\nu}{d\mu} = \frac{d\nu}{dP}\frac{dP}{d\mu}.$$

We will look for a set $F \in \mathcal{A}$ which has the property that $\nu_F[\cdot] := \nu[\cdot \cap F]$ is σ-finite and $\nu[A \cap F^c] = \infty P[A \cap F^c]$. Once we have such a set we can use Theorem 168 to construct $\frac{d\nu_F}{dP}$ and define $\frac{d\nu}{dP} := \frac{d\nu_F}{dP} + \infty I_{F^c}$. This $\frac{d\nu}{dP}$ has the requred properties since

$$\begin{aligned}
\int_A \Big[\frac{d\nu_F}{dP} + \infty I_{F^c}\Big] dP &= \int_A \frac{d\nu_F}{dP} dP + \int_A \infty I_{F^c} dP \\
&= \nu_F[A] + \infty P[A \cap F^c] \\
&= \nu[A \cap F] + \nu[A \cap F^c] \\
&= \nu[A].
\end{aligned}$$

To construct such an F we find the biggest "σ-finite set" as follows

$$\begin{aligned}
&\mathcal{F} := \{\cup_{i=1}^\infty A_i \colon A_i \in \mathcal{A} \text{ and } \nu[A_i] < \infty \text{ for all } i\ \} \\
&m := \sup\{P[F] : F \in \mathcal{F}\}.
\end{aligned}$$

The F we want to construct is the one that attains the above supermum.

To find it let $F_n \in \mathcal{F}$ such that $P[F_n] \to m$ and define $F := \cup_{n=1}^\infty F_n$. Now since $\mathcal{F}$ is clearly closed under countable union (since the countable union of a countable unions is again a countable union) we have $F \in \mathcal{F}$ so that

$$m \leftarrow P[F_n] \le P[F] \le m$$

which implies $m = P[F]$.

Now lets see that F has the desired properties. We can immediately see that $\nu[\cdot \cap F]$ is σ-finite using the cover $F^c, A_1, A_2, \ldots$ where the A_i's come from the fact that $F \in \mathcal{F}$ so that $F = \cup_{i=1}^\infty A_i$ for $\nu[A_i] < \infty$. To finish we just need to show, $\nu[A \cap F^c] = \infty P[A \cap F^c]$, which is equivalent to the following equalities

$$P[A \cap F^c] = 0 \Longrightarrow \nu[A \cap F^c] = 0 \tag{79}$$
$$P[A \cap F^c] > 0 \Longrightarrow \nu[A \cap F^c] = \infty. \tag{80}$$

Equation (79) follows directly from the fact that $\nu \ll P$. We show (80) by contradiction. Assume $P[A \cap F^c] > 0$ but also $\nu[A \cap F^c] < \infty$. Therefore $(A \cap F^c) \cup F \in \mathcal{F}$. But this contradict the maximal property of $P[F]$ as follows

$$m = P[F] < P[A \cap F^c] + P[F] = P[(A \cap F^c) \cup F] \le m.$$

Note, the above inequality is where we use the fact that P is a probability measure. This is a contradiction and therefore (80) holds. □

Theorem 170 (Properties of Radon-Nikodym derivatives). *Suppose $\mu, \sigma, \nu, \nu_1, \nu_2, \ldots$ are measures on $(\Omega, \mathcal{A})$.*

1. *Suppose $\nu_1, \nu_2 \lll \mu$ for $c_1, c_2 \ge 0$. Then $c_1\nu_1 + c_2\nu_2 \ll \mu$ and*
$$\frac{d(c_1\nu_1 + c_2\nu_2)}{d\mu} = c_1\frac{d\nu_1}{d\mu} + c_2\frac{d\nu_2}{d\mu} \quad \mu\text{-a.e.}$$
2. *Assume $\nu_1, \nu_2 \lll \mu$. Then*
$$\nu_1 \le \nu_2 \text{ if and only if } \frac{d\nu_1}{d\mu} \le \frac{d\nu_2}{d\mu} \ \mu\text{-a.e.}$$
3. *If $\nu_n[A]$ is non-decreasing for each $A \in \mathcal{A}$ and $\nu_n \lll \mu$ then*
$$\frac{d\nu_n}{d\mu} \xrightarrow{\mu\text{-a.e.}} \frac{d\nu}{d\mu}$$
where $\nu[A] := \lim_n \nu_n[A]$.
4. *Assume $\nu \lll \mu$. Then*
$$\nu \text{ is finite if and only if } \frac{d\nu}{d\mu} \text{ is } \mu\text{-integrable}$$
5. *Assume $\nu \lll \mu$. Then*
$$\nu \text{ is } \sigma\text{-finite if and only if } \frac{d\nu}{d\mu} < \infty \quad \mu\text{-a.e.}$$

6. *If $\nu \lll \sigma$, $\sigma \lll \mu$ and μ is σ-finite then $\nu \ll \mu$ and*

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

μ-a.e.

7. *If $\mu, \nu \lll \sigma$ then*

$$\frac{d\nu}{d\mu} = \begin{cases} \frac{d\nu}{d\sigma}/\frac{d\mu}{d\sigma} & \text{on } \{\omega\colon \frac{d\mu}{d\sigma}(\omega) > 0\} \\ 0 & \text{otherwise} \end{cases}$$

μ-a.e.

8. *If $\mu \lll \nu$ and $\nu \lll \mu$ then $\frac{d\nu}{d\mu} > 0$ μ-a.e. and*

$$\frac{d\mu}{d\nu} = \frac{1}{d\nu/d\mu}$$

ν-a.e.

Proof. See Exercise 53. □

Theorem 171 (**Lebesgue decomposition).** *Let P and Q be two probability measures on $(\Omega, \mathcal{A})$. There exists a P-null set N and a function $\delta \in \mathscr{N}$ such that*

$$Q[A] = \int_A \delta dP + Q[A \cap N] =: Q_a[A] + Q_s[A]$$

for all $A \in \mathcal{A}$. N is Q-unique, δ is P-unique and $Q = Q_a + Q_s$ is the unique decomposition of a absolutly continuous measure with respect to P and a singluar measure with respect to P. Moreover, δ is the P-largest $\delta \in \mathscr{N}$ such that $\int_A \delta dP \le Q[A]$ for all $A \in \mathcal{A}$.

Definition 80 (**Signed measure).** *If $\mathcal{A}$ is a σ-field of Ω-sets, then $\mu : A \to \overline{\mathbb{R}}$ is a* **signed measure** *if $\mu(\varnothing) = 0$ and $\mu\big(\bigcup_{k=1}^{\infty} A_k\big) = \sum_{k=1}^{\infty} \mu(A_k)$ for all disjoint $A_1, A_2, \ldots \in \mathcal{A}$.*

Theorem 172 (**Hahn-Jordan decomposition).** *If μ is a signed measure on $(\Omega, \mathcal{A})$ then one can write Ω as the disjoint unions of sets S^+ and S^- such that*

- *$\mu(A) \ge 0$ for all $\mathcal{A}$-sets $A \subset S^+$*
- *$\mu(A) \le 0$ for all $\mathcal{A}$-sets $A \subset S^-$.*

Moreover μ has the following decomposition

$$\mu[A] = \mu^+[A] - \mu^-[A]$$

where $\mu^{\pm}[B] = \sup\{\pm\mu[A]\colon A \subset B, A \in \mathcal{A}\}$.

Proof. Since μ can not take on both the values $-\infty$ and ∞ we can suppose wlog that μ does not assume the value $-\infty$.

(*Closure properties of strictly negative sets*) A set $S \in \mathcal{A}$ is said to be **strictly negative** if $\mu[A] \le 0$ for all $\mathcal{A}$-sets $A \subset S$. Notice first the somewhat trivial fact that every subset of a strictly negative set is also strictly negative. We also show that the set of all strictly negative sets is closed under countable union. To see why let $S_1, S_2 \ldots$ denote strictly negative sets. Set $S := \cup_{i=1}^{\infty} S_i = \cup_{i=1}^{\infty} S_i^*$ where $S_i^* := S_i - (S_1 \cup \cdots \cup S_{i-1})$ and show that S is strictly negative. Since S_i is strictly negative so is S_i^* (since it is a subset of S_i). Let $A \subset S$ be $\mathcal{A}$-measurable. Since the S_i^* are disjoint we have that

$$\mu[A] = \mu[A \cap S] = \textstyle\sum_{i=1}^{\infty} \underbrace{\mu[A \cap S_i^*]}_{\le 0} \le 0.$$

Therefore S is strictly negative as was to be shown.

(*Construct S^-*) We will define S^- to be a set which attains the following infimum

$$m := \inf\{\mu[S]\colon S \text{ is a strictly negative } \mathcal{A}\text{-set}\}.$$

To see that such a set exists let S_n be strictly negative sets such that $\mu[S_n] \to m$. Put $S^- := \cup_{n=1}^{\infty} S_n$. Then S^- and hence $S^- - S_n$ are both strictly negative by the closure properties of such sets. Therefore

$$\begin{aligned} m \le \mu[S^-] &= \mu[S_n \cup (S^- - S_n)] \\ &= \mu[S_n] + \mu[S^- - S_n] \\ &\le \mu[S_n] \to m. \end{aligned}$$

Therefore $\mu[S^-] = m$ and hence the infimum is attained.

(*Extraction lemma*) The extraction lemma says

> *For any $\mathcal{A}$-set A there exists a strictly negative $\mathcal{A}$-set $N \subset A$ such that $\mu[N] \le \mu[A]$.*

This lemma is trivial for sets A with positive measure (by choosing $N = \varnothing$). So suppose $\mu[A] < 0$. The idea is to repeatedly extract as much positive measure as possible from A, the remainder should strictly negative. Set

$$\mu^+[A] := \sup\{\mu[B]\colon B \subset A, B \in \mathcal{A}\} \tag{81}$$

and recursively choose $\mathcal{A}$-sets B_n such that

$$\underbrace{B_n \subset A - (B_1 \cup \cdots \cup B_{n-1})}_{\text{remove some of what is left}} \tag{82}$$

$$\underbrace{\mu[B_n] \ge \tfrac{1}{2}\mu^+[A - (B_1 \cup \cdots \cup B_{n-1})] \wedge 1}_{\text{make sure it has some positive measure}}. \tag{83}$$

Note: the '$\wedge$' is there just to avoid having to deal with measure $= \infty$. Also the 1/2 is in front of μ^+ since we don't know that the *sup* in (81) is attained. Now removing $\cup_{n=1}^{\infty} B_n$ from A should give us a strictly negative set. Define our candidate for the strictly negative set $N := A - \cup_{n=1}^{\infty} B_n$. We just need to show $\mu^+[N] \le 0$

Notice that $\mu[A] = \mu[\cup_{n=1}^{\infty} B_n] + \mu[N]$. Remember that $-\infty < \mu[A] < 0$ by the assumption that $\mu[A] < 0$ and that μ doesn't take the value $-\infty$. Therefore we have that $\mu[\cup_{n=1}^{\infty} B_n] =$

$\sum_{n=1}^{\infty} \mu[B_n] < \infty$ (we are using the fact that we picked the B_n's to be disjoint). Therefore $\mu[B_n] \to 0$. This is key. First it says that for all large n we can ignore the '$\wedge$' in (83) and have that $\mu[B_n] \geq \frac{1}{2}\mu^+[A - (B_1 \cup \cdots \cup B_{n-1})]$. Second we use $\mu[B_n] \to 0$ to bound $\mu^+[N]$ as follows

$$\begin{aligned} \mu^+[N] &= \mu^+[A - \cup_{n=1}^{\infty} B_n] \\ &\leq \mu^+[A - \cup_{n=1}^{m} B_n], \text{ larger sup set} \\ &\leq 2\mu[B_m], \text{ discussed above} \\ &\to 0. \end{aligned}$$

Therefore $\mu^+[N] \leq 0$ and hence N is strictly negative.

(*Show S^+ has the right properties*) Define $S^+ := (S^-)^c$ and lets show it is strictly positive. Let $A \subset S^+$ and $A \in \mathcal{A}$. We need to show $\mu[A] \geq 0$. Extract a strictly negative $N \subset A \subset S^+$ such that $\mu[N] \leq \mu[A]$. Notice that $N \cap S^- = \varnothing$ and $N \cup S^-$ is strictly negative. Therefore

$$m \leq \mu[N \cup S^-] = \mu[N] + \mu[S^-] \leq \mu[A] + m$$

which shows that $\mu[A]$ must be positive as was to be shown (note $\mu[N] \leq \mu[A]$ comes from extraction). Therefore $\mu[A] \geq 0$ as was to be shown.

(*Jordan decomposition*) Notice that

$$\mu[A] = \mu[A \cap S^+] + \mu[A \cap S^-] =: \mu^+[A] + \mu^-[A].$$

We need to show

$$\begin{aligned} \mu[A \cap S^+] &= \sup\{\ \ \mu[B] \colon B \subset A, B \in \mathcal{A}\} \\ -\mu[A \cap S^-] &= \sup\{-\mu[B] \colon B \subset A, B \in \mathcal{A}\}. \end{aligned}$$

Since $B \subset A$ it will be sufficient to conclude that

$$\begin{aligned} \mu[B] &= \mu[B \cap S^+] + \mu[B \cap S^-] \\ &\leq \mu[A \cap S^+] + \text{negative} \end{aligned}$$

and

$$\begin{aligned} -\mu[B] &= -\mu[B \cap S^+] - \mu[B \cap S^-] \\ &\leq \quad \text{negative } - \mu[A \cap S^-]. \end{aligned}$$

where the only thing we need is that $\mu[B \cap S^+] \leq \mu[A \cap S^+]$ and $\mu[B \cap S^-] \geq \mu[A \cap S^-]$. But these two inequalities are easy since $\mu[A \cap S^+] = \mu[(A - B) \cap S^+] + \mu[B \cap S^+] \geq \mu[B \cap S^+]$ and $\mu[A \cap S^-] = \mu[(A - B) \cap S^-] + \mu[B \cap S^-] \leq \mu[B \cap S^-]$. $\square$

Exercise 52. *Referring to the proof of Theorem 168 show that f_P and f_Q are both continuous linear functionals over $L^2(W)$.*

Exercise 53. *Prove Theorem 170.*

17.1 Application: dP/dQ for random fields

18 Conditional expectation

We start with the definition of the expected value of a random variable X with respect to a sub-σ-field $\mathcal{B}$, denoted $E^{\mathcal{A}}X$. The basic idea is to define $E^{\mathcal{A}}X$ as a Radon-Nikodym derivative. We then use this to construct $E(X|Y)$ and $E(X|Y=y)$. Play close attention to the fact $E(X|Y=y)$ is only unique up to a modification on PY^{-1}-null sets in y. After we define conditional probability distributions, namely $\mathcal{L}_{X|Y=y}$, we will show that once can construct a special version of $E(X|Y=y)$ that has nice properties.

Remark: Previously in the notes we have used notation such as X or Y to denote random variables (or vectors). In particular, measurable maps defined on probability space which map into $\mathbb{R}$ (or $\mathbb{R}^d$). In this section we will slightly depart from that notational convention and generally write X and Y, etc. for extended random variables.

18.1 Definition of $E^{\mathcal{A}}(X)$

Theorem 173 (Construction of $E^{\mathcal{A}}X$). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X be a P-quasi-integrable extended random variable on Ω. Let $\mathcal{A} \subset \mathcal{F}$ be a sub σ-field. Then there exists a $\mathcal{A}$-measurable, P-quasi-integrable extended random variable $E^{\mathcal{A}}(X)$ such that*

$$\int_A X\,dP = \int_A E^{\mathcal{A}}(X)\,dP \quad \text{for all } A \in \mathcal{A}. \tag{84}$$

Moreover $E^{\mathcal{A}}X$ is P-unique.

Proof. My game plan is to show the result for non-negative random variables then define the general case with $E^{\mathcal{A}}(X^+) - E^{\mathcal{A}}(X^-)$.

Start by assuming $X \geq 0$. Notice that

$$\nu_{\mathcal{A}}(A) := \int_A X\,dP$$

is a measure over $\mathcal{A}$. Let $P_{\mathcal{A}}$ denote the restriction of P to the sub σ-field $\mathcal{A}$. Now we clearly have $\nu_{\mathcal{A}} \lll P_{\mathcal{A}}$ over $\mathcal{A}$. The Radon-Nikodym Theorem 169 gives that $d\nu_{\mathcal{A}}/dP_{\mathcal{A}}$ exists and is $\mathcal{A}$-measurable and P-quasi-integrable. Now

$$E^{\mathcal{A}}(X) := \frac{d\nu_{\mathcal{A}}}{dP_{\mathcal{A}}}$$

has all the required properties. In particular $E^{\mathcal{A}}X$ is $\mathcal{A}$-measurable, P-quasi-integrable and for all $A \in \mathcal{A}$ we have

$$\int_A X dP = \nu_{\mathcal{A}}(A) = \int_A E^{\mathcal{A}}(X) dP_{\mathcal{A}} = \int_A E^{\mathcal{A}}(X) dP$$

where the last equation follows by a change-of-variables Theorem 74 (setting T to be the identity map).

To extend to all P-quasi-integrable random variables X we need to ensure that $E^{\mathcal{A}}(X^+) - E^{\mathcal{A}}(X^-)$ is defined when X is P-quasi-integrable. To this end suppose *wlog* that $E(X^+) < \infty$. We show $E^{\mathcal{A}}(X^+) < \infty$. In this case $\nu_{\mathcal{A}} := \int_\bullet X^+\,dP$ is a finite measure which implies (by item 4 in Theorem 170) that $E^{\mathcal{A}}(X^+) := \frac{d\nu_{\mathcal{A}}}{dP_{\mathcal{A}}}$ is P-integrable and therefore finite P-a.e. Therefore, we may, and do, change $E^{\mathcal{A}}(X^+)$ on a P-null set, without destroying condition (84), so that $E^{\mathcal{A}}(X^+) < \infty$ which ensures $E^{\mathcal{A}}(X^+) - E^{\mathcal{A}}(X^-)$ is defined and, since $E^{\mathcal{A}}(X^+) \in L_1(P)$, $E^{\mathcal{A}}(X^+) - E^{\mathcal{A}}(X^-)$ is quasi-integrable and has the right integration properties.

Uniqueness follows directly from the *uniqueness of densities* Theorem 12 since any $\mathcal{A}$-measurable, P-quasi-integrable $E^{\mathcal{A}}X$ which satisfies the right-hand-side of (84) is P-unique (to apply that theorem I'm using the fact that the base measure P is σ-finite). □

Example 3 (Smoothing property of $E^{\mathcal{A}}X$). *For example work with $([0,1), \mathcal{B}^{[0,1)}, \mathcal{L})$. Let $\mathcal{A} := \{[0,1), \varnothing, [0,1/2), [1/2,1)\}$. Show that*

$$[E^{\mathcal{A}}X](\omega) = \begin{cases} \text{average of } X \text{ over } [0,1/2) & \text{if } \omega \in [0,1/2) \\ \text{average of } X \text{ over } [1/2,1) & \text{if } \omega \in [1/2,1). \end{cases}$$

Example 4 (Resolution of $E^{\mathcal{A}}X$ as expressing information). *A nice heuristic for understanding how $\left(E^{\mathcal{A}}X\right)(\omega)$ is expressing partial information is that one can think of $\left(E^{\mathcal{A}}X\right)(\omega)$ as the average value of X (wrt measure P) over the smallest event in $\mathcal{A}$ containing ω (however, this only holds rigorously when the σ-field is generated by a countable partition of Ω). The smaller the smallest event it is, the more information/resolution you have.*

In particular let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \mathcal{F}$ be an increasing sequence of sub σ-fields where lets set $\mathcal{F}_0 := \{\Omega, \varnothing\}$. Then the corresponding conditional expected values has increasing resolution from no resolution at all, i.e. $E(X)$, to full resolution, i.e. X

$$\begin{aligned} &E^{\mathcal{F}_0}(X) = E(X) \\ &E^{\mathcal{F}_1}(X) \\ &\big\downarrow \text{ increasing resolution} \\ &E^{\mathcal{F}}(X) = X. \end{aligned}$$

Example 5 (Viewing $E^{\mathcal{A}}(X)$ as a projection). *Another way to look at $E^{\mathcal{A}}(X)$ is with projection. This only works when $X \in L^2(P)$. Let S denote the subset of $L_2(P)$ which are $\mathcal{A}$-measurable. It's easy to see that S is a closed linear subspace of $L_2(P)$. Then we can project X onto S, denoted $\mathcal{P}_S X$, which has the property*

$$(X - \mathcal{P}_S X) \perp W$$

for all $W \in S$. Therefore $E[(X - \mathcal{P}_S X)W] = 0$ for all $W \in S$. Therefore $E[XW] = E[(\mathcal{P}_S X)W$ for all $W \in S$. Substatuting $W = I_A$ in the last equation for some $A \in \mathcal{A}$ gives

$$\int_A X dP = \int_A \mathcal{P}_S X dP.$$

This shows that $\mathcal{P}_S X$ serves as $E^{\mathcal{A}}X$.

Theorem 174 (**Smoothing properties of $E^{\mathcal{A}}$**). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{A}_1, \mathcal{A}_2$ be sub σ-fields of $\mathcal{F}$. Suppose that Y, X are P-quasi-integrable extended random variables on $(\Omega, \mathcal{F}, P)$. Then*

1. $E(E^{\mathcal{A}}X) =_{a.e.} E(X)$
2. *If $\mathcal{A}_1 \subset \mathcal{A}_2$ then $E^{\mathcal{A}_1}(E^{\mathcal{A}_2}X) =_{a.e.} E^{\mathcal{A}_2}(E^{\mathcal{A}_1}X) =_{a.e.} E^{\mathcal{A}_1}X$.*
3. $E^{\mathcal{A}}X \in Q^{\pm}(P) \iff X \in Q^{\pm}(P)$
4. *If $XY \in Q(P)$ and X is $\mathcal{A}$-measurable (but not necessarily in $Q(P)$) then $E^{\mathcal{A}}(XY) =_{a.e} XE^{\mathcal{A}}Y$.*

Theorem 175 (**Expected value properties of $E^{\mathcal{A}}$**). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{A}$ be a sub σ-field of $\mathcal{F}$. Suppose that Y, X, X_1, X_2 are extended random variables on $(\Omega, \mathcal{F}, P)$. Then*

1. **Monotonicity:** *If $X, Y \in Q(P)$ then*

$$X \leq Y \ P\text{-a.e.} \Longrightarrow E^{\mathcal{A}}(X) \leq E^{\mathcal{A}}(Y) \ P\text{-a.e.}$$

2. **Linearity:** *If $X \in Q(P)$ and $\alpha \in \mathbb{R}$ or $X \in \mathscr{N}$ and $\alpha \in \{\infty, -\infty\}$ then $\alpha X \in Q(P)$*

$$E^{\mathcal{A}}(\alpha X) =_{a.e.} \alpha E^{\mathcal{A}}(X)$$

If $X, Y, X+Y \in Q(P)$ then

$$I_A E^{\mathcal{A}}(X+Y) =_{a.e.} I_A E^{\mathcal{A}}(X) + I_A E^{\mathcal{A}}(Y)$$

where $A := \{E^{\mathcal{A}}(X) + E^{\mathcal{A}}(Y) \neq \pm\infty \mp \infty\}$.

3. **Continuous from below***:*

$$0 \leq X_n \uparrow X \ P\text{-a.e.} \Longrightarrow E^{\mathcal{A}}(X_n) \uparrow E^{\mathcal{A}}(X) \ P\text{-a.e.}$$

4. **Fatou:** *If $X_n \geq 0$ a.e. then*

$$E^{\mathcal{A}}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} E^{\mathcal{A}}(X_n) \ P\text{-a.e.}$$

5. **DCT:** *If $X_n, X \in Q(P)$ and $X_n \xrightarrow{ae} X$ then*

$$\lim_{n\to\infty} E^{\mathcal{A}}(X_n) = E^{\mathcal{A}}(X) \ P\text{-a.e. on } \{E^{\mathcal{A}}(\sup_n |X_n|) < \infty\}.$$

18.2 Defining $E(X|Y)$ and $E(X|Y = y)$

Definition 81. *Suppose X is an extended random variable on $(\Omega, \mathcal{F}, P)$. Suppose $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ is another measurable space and $Y : \Omega \to \mathcal{Y}$ which is $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ measurable. Then*

$$E(X|Y) := E^{\sigma\langle Y\rangle}X$$

Corollary 23 ($E(X|Y)$ **is a function of** Y). *There exists a $\mathcal{F}^{\mathcal{Y}}$-measurable function $g \colon \mathcal{Y} \to \bar{\mathbb{R}}$ such that $E(X|Y)(\omega) = g(Y(\omega))$ for all $\omega \in \Omega$.*

Proof. This follows directly from Corollary 7 since, by definition, $E(X|Y) = E^{\sigma\langle Y\rangle}X$ is measurable with respect to $\sigma\langle Y\rangle$. □

At times we use the notation $E(X|Y = y)$, for $y \in \mathcal{Y}$, to denote $g(y)$ where $E(X|Y)(\omega) = g(Y(\omega))$. However, since g can be modified on P on PY^{-1} null sets of y's is not meaningful to talk about $E(X|Y = y)$ at a fixed y, but rather about how $E(X|Y = y)$ integrates over y.

Corollary 24 (**Some obvious properties**). *Let X be an extended random variable on the probabilist space $(\Omega, \mathcal{F}, P)$. Let $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be a measure space. Let $Y : \Omega \to \mathcal{Y}$ be $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ measurable. Then*

1. $E(X) =_{a.e.} E(E(X|Y))$*;*
2. *$E(f(Y)X|Y)) =_{a.e.} f(Y)E(X|Y)$ whenever $f(y) : \mathcal{Y} \to \bar{\mathbb{R}}$ is $\mathcal{F}^{\mathcal{Y}}$-measurable and $Xf(Y) \in Q(P)$.*

Exercise 54. *Let $\mathcal{P}$ be a π-system generating the sub-σ-field $\mathcal{A}$ of $\mathcal{F}$ such that $\Omega \in \mathcal{P}$, and let $X : \Omega \to [0, \infty]$ be $\mathcal{F}$-measurable and P-integrable function. Suppose also that Y is $\mathcal{A}$-measurable P-integrable function such that $\int_A X\,dP = \int_A Y\,dP$ for all $A \in \mathcal{P}$. Show that Y is a version of $E^{\mathcal{A}}X$.*

Exercise 55. *Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be sub-σ-fields of $\mathcal{F}$ and for $i = 1, 2$ let $\mathcal{X}_i$ be the collection of $\mathcal{A}_i$-measurable mappings from Ω to $[0, \infty]$. Notice that $\mathcal{A}_1$ and $\mathcal{A}_2$ are* **independent***, written $\mathcal{A}_1 \perp \mathcal{A}_2$ if and only if $E(X_1X_2) = E(X_1)E(X_2)$ for all $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$. $\mathcal{A}_1$ and $\mathcal{A}_2$ are said to be* **conditionally independent** *given a sub-σ-field $\mathcal{B}$ of $\mathcal{F}$, written $\mathcal{A}_1 \perp_{\mathcal{B}} \mathcal{A}_2$, if and only if*

$$E^{\mathcal{B}}(X_1X_2) = E^{\mathcal{B}}(X_1)E^{\mathcal{B}}(X_2) \text{ for all } X_1 \in \mathcal{X}_1,\ X_2 \in \mathcal{X}_2.$$

Also let $\mathcal{C} \vee \mathcal{D}$ denote $\sigma\langle \mathcal{C}, \mathcal{D}\rangle$ when $\mathcal{C}$ and $\mathcal{D}$ are two collections of events on Ω. Show that

1. *$\mathcal{A}_1 \perp_{\mathcal{B}} \mathcal{A}_2 \iff E^{\mathcal{B}}(I_{A_1}I_{A_2}) = E^{\mathcal{B}}(I_{A_1})E^{\mathcal{B}}(I_{A_2})$ a.e. for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.*
2. *$\mathcal{A}_1 \perp_{\mathcal{B}} \mathcal{A}_2 \iff E^{\mathcal{A}_1 \vee \mathcal{B}}X_2 = E^{\mathcal{B}}X_2$ a.e. for all $X_2 \in \mathcal{X}_2$.*
3. *$\mathcal{A}_1 \perp \mathcal{A}_2 \iff E^{\mathcal{A}_1}X_2 = EX_2$ a.e. for all $X_2 \in \mathcal{X}_2$.*
4. *$\mathcal{A}_1 \perp_{\mathcal{B}} \mathcal{A}_2 \iff (\mathcal{A}_1 \vee \mathcal{B}) \perp_{\mathcal{B}} (\mathcal{A}_2 \vee \mathcal{B})$.*
5. *$(\mathcal{A}_1 \vee \mathcal{B}) \perp \mathcal{A}_2 \iff \mathcal{B} \perp \mathcal{A}_2$ and $\mathcal{A}_1 \perp_{\mathcal{B}} \mathcal{A}_2$.*

Hint for 2. ($\Longrightarrow$), it suffices (why?) to consider the case where X_2 is integrable; apply the preceding exercise with $X = X_2$ and $\mathcal{P} = \{A_1 \cap B \colon A_1 \in \mathcal{A}_1 \text{ and } B \in \mathcal{B}\}$.

18.3 The substitution fallacy

The substitution fallacy:

> *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be two other measurable spaces. Let $X \colon \Omega \to \mathcal{X}$ and $Y \colon \Omega \to \mathcal{Y}$ be measurable maps into their respective measurable spaces. Let $f(x, y) : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be*

$\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$-measurable and quasi-integrable with respect to $P(X,Y)^{-1}$. Then

$$E(f(X,Y)|Y=y) = E(f(X,y)|Y=y) \qquad (85)$$

The problem with the above statement is that it is not clear what is meant by equation (85). The left hand side is $g(y)$ where $g(Y) = E(f(X,Y)|Y) = E^{\sigma\langle Y\rangle} f(X,Y)$ is a function of Ω. Now what do we mean by the right hand side of (85)? If we fix y, maybe we consider $f(X,y)$ to be a function on Ω. Then there exists $g_y : \mathcal{Y} \to \bar{\mathbb{R}}$ such that $g_y(Y) = E(f(X,y)|Y)$. In this case we are asking if $g(y) = g_y(y)$. One can immediately see the problem. The functions g_y is PY^{-1} unique. So if, say, that $P(Y=y) = 0$ for every y, then I can change g_y at y to be any number I want and not destroy the fact that $g_y(Y)$ would serve as a version of $E(f(X,y)|Y)$. This implies that $g_y(y)$ is not well defined.

Theorem 176 (A correct version of substitution). *If, in addition to the antecedent presented in the substitution fallacy, X and Y are independent, then $E[f(X,y)]$ serves as a version of $E(f(X,Y)|Y=y)$.*

A complete resolution of the substitution fallacy can not be resolved until the next section when we talk about regular conditional probability distributions

19 Conditional probability

The main story is that $P(A|B)$ doesn't have any meaning when $P(B) = 0$. We can only make sense of this when B has the form $Y = y$ for some y. However, there may be multiple different random variables, say Y_1 and Y_2, such that $\{Y_1 = y_1\} = \{Y_2 = y_2\} = B$ but $P(A|Y_1 = y_1) \neq P(A|Y_2 = y_2)$. So, in effect, what we mean by $P(A|B)$ depends on what which random quantity Y we choose.

Definition 82 (**Probability of A given Y or $\mathcal{A}$**). *Let $(\Omega, \mathcal{F}, P)$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be a probability space and measure space, respectively. Let $Y : \Omega \to \mathcal{Y}$ be ⓜ$\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ and $\mathcal{A} \subset \mathcal{F}$ be a sub-σ-field. Let $F \in \mathcal{F}$ and define*

$$
\begin{aligned}
P(F|\mathcal{A}) &:= E^{\mathcal{A}}(I_F) \\
P(F|Y) &:= E(I_F|Y) \\
P(F|Y = \bullet) &:= E(I_F|Y = \bullet)
\end{aligned}
$$

Let's step back and unwind the definition a bit. First notice that by the definition of $E(I_F|Y)$ we have that

$$P\big(A \cap F\big) = \int_A P\big(F|Y\big)dP$$

for all $A \in \sigma\langle Y\rangle$. Since every $A \in \sigma\langle Y\rangle = Y^{-1}(\mathcal{F}^{\mathcal{Y}})$ has the form $Y^{-1}(B)$ for some $B \in \mathcal{F}^{\mathcal{Y}}$, one can write $A = \{Y \in B\}$ and therefore

$$
\begin{aligned}
P\big(\{Y \in B\} \cap F\big) &= \int_{\{Y\in B\}} P(F|Y)dP \\
&= \int_\Omega I_{\{Y\in B\}} P(F|Y)dP \\
&= \int_\Omega I_B(y)P(F|Y = y)dPY^{-1}(y) \\
&= \int_B P(F|Y = y)dPY^{-1}(y)
\end{aligned}
$$

which is what one might call the law of total probability. Notice that in the case that F corresponds to the event $\{X \in A\}$ for some extended random variable on $(\Omega, \mathcal{F}, P)$, then the above equation simplifies to

$$P\big(Y \in B, X \in A\big) = \int_B P\big(X \in A|Y = y\big)dPY^{-1}(y)$$

At this point it would seem like there is nothing else to do. I've defined the conditional probability of F or $\{X \in A\}$ given $Y = y$. But, notice we still don't know that if we fix y, that $P(F|Y = y)$ is a genuine probability measure when we let F vary over $\mathcal{F}$. The problem is that if we set $P(F|Y) = g_F(Y)$, then we may not necessarily have that $g_F(y)$ is a probability measure on $(\Omega, \mathcal{F})$ for each fixed y. The problem is that I'm free to change $g_F(y)$ on a PY^{-1} null set of y's for each fixed F. What we want is to find a version of $g_F(y)$ for each $F \in \mathcal{F}$ that satisfies the following definition.

Definition 83 (**Conditional probability distribution**). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be a measurable space and $Y : \Omega \to \mathcal{Y}$ be ⓜ$\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$. A map $g(F|y) : \mathcal{F} \times \mathcal{Y} \to [0, 1]$ is a* **conditional probability distribution on** $(\Omega, \mathcal{F})$ **given** Y*, if*

1. *For every $y \in \mathcal{Y}$, the function $g(\bullet|y)$ is a probability measure on $\mathcal{F}$;*
2. *For every $F \in \mathcal{F}$, the function $g(F|\bullet)$ is ⓜ$\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$, PY^{-1}-quasi-integrable and*

   $$P(A \cap F) = \int_A g(F|y)dPY^{-1}(y)$$

   for all $A \in \sigma\langle Y\rangle$.

We will mainly be interested in conditional probability distributions of a random X with respect to some random Y. This is technically subsumed by the previous definition by setting $\{X \in A\} = F = \mathcal{F}$ but it will be more clear if we define a specific definition.

Definition 84 (**Conditional probability distribution of X given $Y = y$**). *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be two measurable spaces. Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ be two functions which are ⓜ$\mathcal{F}/\mathcal{F}^{\mathcal{X}}$ and ⓜ$\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ respectively. A map $\mathcal{L}_{X|Y=y}(A) : \mathcal{F}^{\mathcal{X}} \times \mathcal{Y} \to [0, 1]$ is a* **conditional probability distribution of X given Y**, *if*

1. *For each $y \in \mathcal{Y}$, the function $\mathcal{L}_{X|Y=y}(\bullet)$ is a probability measure on $\mathcal{F}^{\mathcal{X}}$.*
2. *For each $A \in \mathcal{F}^{\mathcal{X}}$, the function $\mathcal{L}_{X|Y=\bullet}(A)$ is ⓜ$\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$, PY^{-1}-quasi-integrable and*

   $$P\big(Y \in B,\, X \in A\big) = \int_B \mathcal{L}_{X|Y=y}(A)\, dPY^{-1}(y)$$

   for all $B \in \mathcal{F}^{\mathcal{Y}}$.

So, the main result of this section is to figure out conditions that allow us to construct a version of $E(I_F|Y = y)$, for each F, which forms *conditional probability distribution.* Before we get to existence theorems lets sharpen our intuition on some other theorems (non-existence, uniqueness, the density case, the independence case, the infinite dimensional case)

How to think about a conditional probability distribution: Remember that it doesn't really mean much to talk about $E(X|Y = y)$ for a fixed y when $P[Y = y] = 0$, since $E(X|Y)$ is of the form $g(Y)$ and one is free to change g on a null set of y's. This will continue to be true even , when later, we will show that there is often a particular choice of $E(X|Y = y)$ which has nice properties (like the substitution principle can be proven). The fact remains $E(X|Y = y)$ is still only meaningful to talk about how it integrates over y, not the value at any particular y.

The same message will hold true for $\mathcal{L}_{X|Y=y}$. At any fixed y when $P[Y = y] = 0$, $\mathcal{L}_{X|Y=y}$ has no real meaning. The example presented in the introduction serves as a perfect example of this. Indeed, if you let $Z := X/Y$ and $W := X - Y$, then the results in the next section show that $\mathcal{L}_{X|Z=z}$ and $\mathcal{L}_{X|W=w}$ both exist as regular conditional probability distributions for X given Z and for X given W, respectively. Now, the events $W = 0$ and $Z = 1$ are exactly the same as subsets of Ω. However, $\mathcal{L}_{X|Z=1}[B] \neq \mathcal{L}_{X|W=0}[B]$. There is no contradiction here, except for the fallacy that $P[X \in B|Z = 1]$ or $P[X \in B|W = 0]$ means anything. Rather we can only really make sense about how $P[X \in B|Z = z]$ integrates as a function of z and similarly for $P[X \in B|W = w]$.

One thing to notice, and this might clear things up a bit, is that

$$\lim_{\epsilon\to 0} P[X \in B|Z \in B_1^\epsilon] \neq \lim_{\epsilon\to 0} P[X \in B|W \in B_0^\epsilon]$$

where $B_x^\epsilon := \{y \in \mathbb{R}\colon |x - y| < \epsilon\}$ is the open ball, around x, of radius ϵ. This is one way to make sense of why $\mathcal{L}_{X|Z=1}[B]$ and $\mathcal{L}X|W = 0[B]$ are different. However, in this example, there is a sense in which $\mathcal{L}_{X|Z=z}$ and $\mathcal{L}_{X|W=w}$ are continuous. Therefore the fact that the limits are different is just a manifestation of the fact that they integrate differently over some z and w regions. See theorem 58 for reference to the continuity result.

19.1 Uniquness, density case, etcetra

Section Assumption. *For the remainder of this section, unless otherwise stated, we make with the following assumptions: Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be two measurable spaces. Let $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ be two functions which are ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{X}}$ and ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ respectively.*

Theorem 177 (**Sometimes a cpd does not exist**)**.**

Theorem 178 (**Factor the joint**)**.** *Let $\mathcal{L}_{X,Y}$ be the joint law of (X, Y) (i.e. $P(X, Y)^{-1}$) and $\mathcal{L}_Y$ be the marginal law of Y (i.e. PY^{-1}). Let $\mathcal{L}_{X|Y=y}$ be a conditional probability distribution of X given Y. Then for any $F \in \mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$ the function $\mathcal{L}_{X|Y=y}(F_y)$ is $\mathcal{L}_Y$-quasi-integrable (where $F_y := \{x \in \mathcal{X}\colon (x, y) \in F)$ and*

$$\mathcal{L}_{X,Y}(F) = \int_{\mathcal{Y}} \mathcal{L}_{X|Y=y}(F_y)\, d\mathcal{L}_Y(y). \tag{86}$$

Proof. Notice that $F_y \in \mathcal{F}^{\mathcal{X}}$ for each $y \in \mathcal{Y}$ by Theorem 97 so that $\mathcal{L}_{X|Y=y}[F_y]$ is well defined. Now the results follows from the following three parts of proof.

(*Part I: $\mathcal{L}_{X|Y=y}[F_y]$ is $\mathcal{L}_Y$-quasi-integrable*) Since $\mathcal{L}_{X|Y=y}[F_y]$ is required to take values in $[0, 1]$ it is sufficient to show $\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$-measurablilty. We use good sets. Define

$$\mathcal{G} := \big\{F \in \mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}\colon\ \mathcal{L}_{X|Y=y}[F_y] \text{ is ⓜ } \mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}\big\}. \tag{87}$$

- The measurable rectangles are in $\mathcal{G}$ since $\mathcal{L}_{X|Y=y}[(B \times A)_y] = I_A(y)\mathcal{L}_{X|Y=y}[B]$ and $\mathcal{L}_{X|Y=y}[B]$ is required to be ⓜ $\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$.
- $\mathcal{G}$ is also closed under complementation. In particular, if $F \in \mathcal{G}$ then

  $$\mathcal{L}_{X|Y=y}[(F^c)_y] = \mathcal{L}_{X|Y=y}[(F_y)^c] = 1 - \mathcal{L}_{X|Y=y}[F_y]$$

  which measurable by the closure theorem.
- Finally notice that $\mathcal{G}$ is closed under disjoint union. In particular, suppose $F_1, F_2, \ldots \in \mathcal{G}$ are all disjoint. Then

  $$\mathcal{L}_{X|Y=y}[(\cup_n F_n)_y] = \mathcal{L}_{X|Y=y}[\cup_n (F_n)_y] = \textstyle\sum_n \mathcal{L}_{X|Y=y}[F_n]$$

  which is measurable by the closure theorem.

The above bullets show that $\mathcal{G}$ is a λ-system which is generated by the π system of measurable rectangles. Therefore

$$\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}} = \sigma\langle\text{rectangles}\rangle = \lambda\langle\text{rectangles}\rangle \subset \mathcal{G}$$

where the second '=' follows from Dynkin's $\pi - \lambda$ theorem and '$\subset$' follows from good sets. Therefore $\mathcal{L}_{X|Y=y}[F_y]$ is ⓜ $\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$ for every $F \in \mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$.

(*Part II: RHS of (86) is a probability measure*) This follows since slicing commutes with set operations, $\mathcal{L}_{X|Y=y}$ is a probability measures for each y, and by properties of the integral $\int_{\mathcal{Y}} \bullet\, d\mathcal{L}_Y(y)$, in particular Theorem 17.

(*Part III: (86) holds on rectangles*) Let $B \in \mathcal{F}^{\mathcal{X}}$ and $A \in \mathcal{F}^{\mathcal{Y}}$ so that $B \times A \in \mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$. Notice that $(B \times A)_y = B$ when $y \in A$ and $(B \times A)_y = \varnothing$ when $y \notin A$. Therefore

$$\begin{aligned}
&\int_{\mathcal{Y}} \mathcal{L}_{X|Y=y}[(A \times B)_y]\, d\mathcal{L}_Y(y)\\
&\qquad = \int_A \mathcal{L}_{X|Y=y}[B]\, d\mathcal{L}_Y(y)\\
&\qquad = P[Y \in A, X \in B], \text{ property of } \mathcal{L}_{X|Y=y}\\
&\qquad = \mathcal{L}_{X,Y}[A \times B].
\end{aligned}$$

Therefore (86) holds on the π-system of measurable rectangles which generates $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$. By uniqueness of probability measures on π-system generators we have (86) holds on all of $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$.

□

Theorem 179 (**Uniqueness of cpd**)**.** *Let $g(F|y)$ and $g^*(F|y)$ be two conditional probability distributions on $(\Omega, \mathcal{F})$ given Y. If $\mathcal{F}$ is countably generated then*

$$P\Big(g(F|Y) = g^*(F|Y) \text{ for all } F \in \mathcal{F}\Big) = 1. \tag{88}$$

Proof. Let $\mathcal{F}_0$ be a countable set of generators such that $\mathcal{F} = \sigma\langle\mathcal{F}_0\rangle$. Notice that we can suppose without loss of generality that $\mathcal{F}_0$ is also a π-system (by closing $\mathcal{F}_0$ under finite intersection, which preserves countability). Since measures are equal if they agree on a π-system generating set we have

$$\begin{aligned}
&\big\{g(F|Y) = g^*(F|Y) \text{ for all } F \in \mathcal{F}\big\}\\
&\qquad = \big\{g(F|Y) = g^*(F|Y) \text{ for all } F \in \mathcal{F}_0\big\}
\end{aligned}$$

$$= \bigcap_{F \in \mathcal{F}_0} \{g(F|Y) = g^*(F|Y)\}. \tag{89}$$

Notice that for each $F \in \mathcal{F}_0$, $g(F|\cdot)$ and $g^*(F|\cdot)$ are ⓜ $\mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$. Therefore $g(F|Y(\cdot))$ and $g^*(F|Y(\cdot))$ is ⓜ $\mathcal{F}/\mathcal{B}^{\bar{\mathbb{R}}}$ (by Theorem 52) and hence $\{\omega \in \Omega \colon g(F|Y(\omega)) - g^*(F|Y(\omega)) = 0\}$ is an $\mathcal{F}$-measurable set. Therefore $\{g(F|Y) = g^*(F|Y) \text{ for all } F \in \mathcal{F}\}$ is an $\mathcal{F}$-measurable set.

Now the theorem follows by noticing that for any fixed $F \in \mathcal{F}$ $g(F|Y)$ and $g^*(F|Y)$ are both a version of $E(I_F|Y)$, which is unique PY^{-1} almost everywhere. Therefore $P\big(g(F|Y) = g^*(F|Y)\big) = 1$ and hence, by (89), (88) follows. □

Theorem 180 (The density case). *Let $\mathcal{L}_{X,Y}$ denote the joint distribution of X and Y on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}^X \otimes \mathcal{F}^Y)$ (see Theorem 64 for existence). Suppose μ and σ are σ-finite measures on $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ respectively. If $f(x,y)$ is a density of $\mathcal{L}_{X,Y}$ with respect to $\mu \otimes \sigma$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}})$ then for each $y \in \mathcal{Y}$ define*

$$f_{X|Y=y}(x) := \begin{cases} f(x,y)/f_Y(y), & \text{when } y \in G \\ f_X(x), & \text{when } y \notin G \end{cases}$$

where

$$f_Y(y) := \int_{\mathcal{X}} f(x,y)\, d\mu(x)$$
$$f_X(x) := \int_{\mathcal{Y}} f(x,y)\, d\sigma(y)$$

and $G := \{y \colon 0 < f_Y(y) < \infty\}$. Then $\mathcal{L}_{X|Y=y}[F] := \int_F f_{X|Y=y}(x)\, d\mu(x)$ defines a cpd of X given Y.

Proof. By Fubini, f_X and f_Y are ⓜ $\mathcal{F}^{\mathcal{X}}$ and ⓜ $\mathcal{F}^{\mathcal{Y}}$, respectively, and can be modified on appropriate null sets to take values in $[0,\infty]$. Therefore $f_{X|Y=y}(x)$ is $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$-measurable and non-negative. For one thing, this immediately gives that $\mathcal{L}_{X|Y=y}$ is well defined. Also notice that $P[Y \in G] = 1$. Indeed Fubini also gives that f_Y and f_X are the marginal densities of Y and X, respectively, so that

$$P[Y \in G^c] = \int_{f_Y=0} f_Y(y) d\sigma + \int_{f_Y=\infty} f_Y(y) d\sigma \leq 1.$$

Obviously $\int_{f_Y=0} f_Y(y) d\sigma = 0$ and $\int_{f_Y=\infty} f_Y(y) d\sigma$ must be zero or else it would violate the above upper bound.

(*Conditions for* $\mathcal{L}_{X|Y=y}[\bullet]$) Since we already know that indefinite integrals of positive measurable functions are measures we just check $\int_{\mathcal{X}} f_{X|Y=y}(x)\, d\mu(x) = 1$, which follows easily by the definition of $f_{X|Y=y}$.

(*Conditions for* $\mathcal{L}_{X|Y=\bullet}[F]$) To prove the necessary conditions for $\mathcal{L}_{X|Y=\bullet}[F]$ notice that Fubini's theorem establishes $\int_F f_{X|Y=\bullet}(x)\, d\mu(x)$ is $\mathcal{F}^{\mathcal{Y}}$-measurable and quasi-integrable. Now notice that for each $A \in \mathcal{F}^{\mathcal{Y}}$ and $B \in \mathcal{F}^{\mathcal{X}}$ we have

$$\begin{aligned}
&P[Y \in A, X \in B] \\
&\quad = P[Y \in A \cap G, X \in B], \quad \text{since } P[Y \in G] = 1 \\
&\quad = \int_{(A \cap G) \times B} f(x,y)\, d\mu \otimes \sigma \\
&\quad = \int_{A \cap G} \left[\int_B f(x,y)\, d\mu(x) \right] d\sigma(y), \quad \text{Fubini} \\
&\quad = \int_{A \cap G} \left[\int_B \frac{f(x,y)}{f_Y(y)}\, d\mu(x) \right] f_Y(y)\, d\sigma(y) \\
&\quad = \int_{A \cap G} \mathcal{L}_{X|Y=y}[B] f_Y(y)\, d\sigma(y) \\
&\quad = \int_{A \cap G} \mathcal{L}_{X|Y=y}[B] dPY^{-1} \\
&\quad = \int_A \mathcal{L}_{X|Y=y}[B] dPY^{-1}
\end{aligned}$$

where the last line follows since $I_G = 1$, PY^{-1}-a.e.. □

Theorem 181 (The independence case). *If X and Y are independent then a conditional probability distribution of X given Y exists and a version of it is given by $\mathcal{L}_{X|Y=y} = PX^{-1}$. Conversely if $\mathcal{L}_{X|Y=y} = Q$ for all $y \in \mathcal{Y}$ for some probability measure Q on $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$, then X and Y are independent and $PX^{-1} = Q$.*

Theorem 182 (π-system tool). *Let $\mathcal{P}$ be a π-system generating $\mathcal{F}^{\mathcal{X}}$. If $\{Q_y\}_{y \in \mathcal{Y}}$ is a collection of probabilty measures on $(\mathcal{X}, \mathcal{F}^X)$ such that $Q_\bullet[F]$ is a version of $P[X \in F|Y = \bullet]$ for each $F \in \mathcal{P}$. Then a conditinoal probability distribution of X given Y exists and a version of it is given by $\mathcal{L}_{X|Y=y} = Q_y$.*

Proof. Define $\mathcal{G}$ to be the set of all $B \in \mathcal{F}^{\mathcal{X}}$ which satsifes: *$Q_\bullet[B]$ is $\mathcal{F}^{\mathcal{Y}}$-measurable and $P[Y \in A, X \in B] = \int_A Q_y[B]\, dPY^{-1}(y)$ for all $A \in \mathcal{F}^{\mathcal{Y}}$.* Notice that $\mathcal{G}$ is a λ-system. By assumption $\mathcal{G}$ contains the π-system $\mathcal{P}$. By good sets $\lambda\langle\mathcal{P}\rangle \subset \mathcal{G}$. Then by Dynkin's $\pi - \lambda$ theorem we have $\lambda\langle\mathcal{P}\rangle = \sigma\langle\mathcal{P}\rangle = \mathcal{F}^{\mathcal{X}} \subset \mathcal{G}$. □

19.2 Existance of $\mathcal{L}_{X|Y=y}$

Section Assumption. *For the remainder of this section we continue with the following assumptions: Let $(\Omega, \mathcal{F}, P)$ is a probability space. Let $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be two measurable spaces. Let $X \colon \Omega \to \mathcal{X}$ and $Y \colon \Omega \to \mathcal{Y}$ be two functions which are ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{X}}$ and ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ respectively.*

Lemma 8 (Cumulative distributions give cpd's). *Suppose $(\mathcal{X}, \mathcal{F}^{\mathcal{X}}) = (\mathcal{R}, \mathcal{B}^{\mathbb{R}})$ and there exists a function $F(\bullet|\bullet) \colon \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ such that*

1. *For each $y \in \mathcal{Y}$, $F(\bullet|y)$ is a cumulative distribution function;*
2. *For each $x \in \mathbb{R}$, $F(x|\bullet)$ is a version of $P[X \leq x|Y = \bullet]$.*

Then a conditional probability distribution of X given Y exists. Moreover any version of $\mathcal{L}_{X|Y=y}$ satisfies

$$\mathcal{L}_{X|Y=y}[X \leq x] = F(x|y)$$

for each $y \in \mathcal{Y}$.

Theorem 183 (Little existence theorem). *If $(\mathcal{X}, \mathcal{F}^{\mathcal{X}}) = (\mathbb{R}, \mathcal{B}^{\mathbb{R}})$ then there exists a conditional probability distribution of X given Y.*

Definition 85 (Isomorphic measure spaces). *Two measurable spaces $(\Omega, \mathcal{F})$ and $(\Omega^*, \mathcal{F}^*)$ are said to be* **isomorphic** *if there exists a one-to-one mapping φ from Ω onto Ω^* such that both φ and φ^{-1} are measurable.*

Definition 86 (Standard Borel space). *A measurable space $(\Omega, \mathcal{F})$ is said to be a* **standard Borel space** *if there exists a Borel set B of $\mathbb{R}$ such that $(B, B \cap \mathcal{B}^{\mathbb{R}})$ is isomorphic to $(\Omega, \mathcal{F})$.*

Theorem 184 (Big existence theorem). *If $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ is a standard Borel space then there exists a conditional probability distribution of X given Y.*

19.3 A special version of $E(X|Y)$ using $\mathcal{L}_{X|Y=y}$.

Section Assumption. *For the remainder of this section we continue with the following assumptions: Let $(\Omega, \mathcal{F}, P)$ is a probability space. Let $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}^{\mathcal{Y}})$ be two measurable spaces. Let $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ be two functions which are ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{X}}$ and ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ respectively. Let $\mathcal{L}_{X,Y}$ denote in induced measures $P(X,Y)^{-1}$ and $\mathcal{L}_Y$, $\mathcal{L}_X$ denote the marginal measures PY^{-1} and PX^{-1}.*

Under the case that there exists a conditional probability distribution of X given Y, $\mathcal{L}_{X|Y=y}$, and that $\mathcal{X} \subset \mathbb{R}$, we can define a special version of $E(X|Y)$ as follows

$$E(X|Y = y) := \int_{\mathcal{X}} x \, d\mathcal{L}_{X|Y=y}(x).$$

The fact that this definition has the correct properties of a conditional expected value follows from the following Law of the Iterated integral which is essentially a Fubini-type theorem with more general joint probability measures on $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$.

Theorem 185 (Law of the Iterated Integral, version 1). *Let $f(x,y)\colon \mathcal{X} \times \mathcal{Y} \to [0,\infty]$ be $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$-measurable. Then $\int_{\mathcal{X}} f(x,y)\, d\mathcal{L}_{X|Y=y}(x)$ is defined for all $y \in \mathcal{Y}$, a measurable function of $y \in \mathcal{Y}$ and is quasi-integrable with respect to $\mathcal{L}_Y$. Moreover,*

$$\int_{\mathcal{X}\times\mathcal{Y}} f d\mathcal{L}_{X,Y} = \int_{\mathcal{Y}} \Big[\int_{\mathcal{X}} f(x,y) d\mathcal{L}_{X|Y=y}(x)\Big] d\mathcal{L}_Y(y). \qquad (90)$$

If f is allowed to take negative values then (90) still holds provided $f \in Q(\mathcal{L}_{X,Y})$; in this case the inner integral on the right hand side of (90) is defined for $\mathcal{L}_Y$-a.e. y.

This theorem is a re-statement of the above LII, but is more specific as to how it allows us to define a special version of conditional expected value which can resolve the substitution fallacy.

Theorem 186 (Law of the Iterated Integral, version 2). *Let $f(x,y)\colon \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}$-measurable and quasi-integrable with respect to $\mathcal{L}_{X,Y}$. Define*

$$h(y) := \int_{\mathcal{X}} f(x,y)\, d\mathcal{L}_{X|Y=y}(x). \qquad (91)$$

Then $h(y)$ is defined $\mathcal{L}_Y$-a.e., is $\mathcal{F}^{\mathcal{Y}}$-measurable and $h(Y)$ is a version of $E\big(f(X,Y)|Y\big)$.

Remember that $\mathcal{L}_{X|Y=y}$ is a probability distribution on $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$ for each $y \in \mathcal{Y}$. Therefore we can interpret $h(y)$ as the expected value of $f(X,y)$ where $X \sim \mathcal{L}_{X|y=y}$. In particular,

$$h(y) =_{a.e.} E(f(X,y)|Y = y).$$

Or another way to put it, we can *define* $E(f(X,y)|Y = y)$ to be $h(y)$ (h being defined $\mathcal{L}_Y$-a.e.). Then under this definition the Law of the Iterated Integral tells us that $h(Y)$ is a version of $E(f(X,Y)|Y)$ which resolves the substitution fallacy since

$$E(f(X,Y)|Y = y) =_{a.e.} h(y) =_{a.e.} E(f(X,y)|Y = y)$$

Corollary 25 (Resolving the substitution fallacy). *Let $f(x,y)\colon \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ be ⓜ $\mathcal{F}^{\mathcal{X}} \otimes \mathcal{F}^{\mathcal{Y}}/\mathcal{B}^{\bar{\mathbb{R}}}$ and quasi integrable with respect to $\mathcal{L}_{X,Y}$. If we define $E(f(X,y)|Y = y)$ to denote the function $h(y)$ as defined in (91) then $E(f(X,y)|Y = y)$ is a version of $E(f(X,Y)|Y = y)$.*

Exercise 56. *Prove the Law of the Iterated Integral for conditional probability distributions. Hint: Mimic the proof of Fubinito. Notice that, in the proof of Fubinito, the key was the identity*

$$P_1 \otimes P_2(F) = \int_{\Omega_1} P_2(F_{\omega_1})\, dP_1(\omega_1).$$

For the proof LII notice that they analogous key formula is

$$\mathcal{L}_{X,Y}(F) = \int_{\mathcal{Y}} \mathcal{L}_{X|Y=y}(F_y)\, d\mathcal{L}_Y(y)$$

which we already proved in the notes.

For the following two exercises suppose $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces. Let $\mathcal{F}^{\mathcal{X}}$ and $\mathcal{F}^{\mathcal{Y}}$ be the Borel σ-fields generated by the respective metrics. Let $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ be two functions which are ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{X}}$ and ⓜ $\mathcal{F}/\mathcal{F}^{\mathcal{Y}}$ respectively.

Definition 87. *A conditional probability distribution $\mathcal{L}_{X|Y=y}$ is said to be* **weakly continuous** *if for every $y \in \mathcal{Y}$*

$$\mathcal{L}_{X|Y=y_n} \rightsquigarrow \mathcal{L}_{X|Y=y}$$

whenever $y_n \to y$.

Exercise 57. *Show that if $P(Y \in G) > 0$ for each non-empty open $G \subset \mathcal{Y}$ then any weakly continuous conditional probability distribution (cpd) for X given Y is completely unique. In particular, show that if $\mathcal{L}_{X|Y=y}$ and $\mathcal{L}^*_{X|Y=y}$ are two weakly continuous cpds then $\mathcal{L}_{X|Y=y} = \mathcal{L}^*_{X|Y=y}$ for all $y \in \mathcal{Y}$.*

Hint: Show that whenever $f\colon \mathcal{X} \to \mathbb{R}$ is continuous and bounded then $\int_{\mathcal{X}} f d\mathcal{L}_{X|Y=y} - \int_{\mathcal{X}} f d\mathcal{L}^*_{X|Y=y}$ is zero for all $y \in \mathcal{Y}$.

Exercise 58. *Suppose $P(Y \in G) > 0$ for each non-empty open $G \subset \mathcal{Y}$. Let $B^{\epsilon}_{y_0} := \{y \in \mathcal{Y} \colon d_Y(y, y_0) < \epsilon\}$ and define*

$$\mathcal{L}_{X|Y \in B^{\epsilon}_{y_0}}(\bullet) := \frac{P[X \in \bullet, Y \in B^{\epsilon}_{y_0}]}{P[Y \in B^{\epsilon}_{y_0}]}$$

which is a probability measure on $(\mathcal{X}, \mathcal{F}^{\mathcal{X}})$. Show that if there exists a weakly continuous cpd $\mathcal{L}_{X|Y=y}$ then for each $y_0 \in \mathcal{Y}$

$$\mathcal{L}_{X|Y \in B^{\epsilon}_{y_0}} \rightsquigarrow \mathcal{L}_{X|Y=y_0}$$

as $\epsilon \to 0$.

Hint: When $f \colon \mathcal{X} \to \mathbb{R}$ is continuous and bounded show that

$$E\big(f|Y \in B^{\epsilon}_{y_0}\big) = \frac{E\big[f(X) I_{B^{\epsilon}_{y_0}}(Y)\big]}{P\big[Y \in B^{\epsilon}_{y_0}\big]} = \int_{B^{\epsilon}_{y_0}} \frac{E\big(f|Y = y\big)}{P\big[Y \in B^{\epsilon}_{y_0}\big]} dPY^{-1}(y).$$

Be sure to be precise about what $E\big(f|Y = y\big)$ denotes.

19.4 Application: L_2 Wasserstein metric, optimal coupling, disintegration

Part V
Martingales

Martingales can be thought of the stochastic analog of monotonic sequences of numbers. In particular, a fundamental feature of a monotonic sequences of numbers is that they always has a limit and that limit is finite when the sequence is bounded. In a way, most of limit theorems in martingale theory have this type of flavor, given some kind of a stochastic bound on a sub-martingale, it is bound to have a limit of some sort.

20 Basic Theory

Section Assumption. *For this section let $(\Omega, \mathcal{F}, P)$ denote an arbitrary probability space.*

Definition 88. *A sequence of sub-σ-fields $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is called a* **filtration** *if $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$.*

Definition 89. *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to be* **adapted to the filtration** *$(\mathcal{F}_n)_{n\in\mathbb{N}}$ if X_n is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}$.*

Definition 90. *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to be* **a martingale with respect to a filtration** *$(\mathcal{F}_n)_{n\in\mathbb{N}}$ if $(X_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, each X_n is integrable and*

$$E^{\mathcal{F}_n}(X_{n+1}) =_{a.e.} X_n, \quad \text{for } n \in \mathbb{N}.$$

Definition 91. *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to be* **a submartingale with respect to a filtration** *$(\mathcal{F}_n)_{n\in\mathbb{N}}$ if $(X_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, each X_n is integrable and*

$$E^{\mathcal{F}_n}(X_{n+1}) \geq_{a.e.} X_n, \quad \text{for } n \in \mathbb{N}.$$

Definition 92. *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to be* **a supermartingale with respect to a filtration** *$(\mathcal{F}_n)_{n\in\mathbb{N}}$ if $(X_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, each X_n is integrable and*

$$E^{\mathcal{F}_n}(X_{n+1}) \leq_{a.e.} X_n, \quad \text{for } n \in \mathbb{N}.$$

Definition 93. *The* **natural filtration** *of a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is $\mathcal{F}_n := \sigma\langle X_1, X_2, \ldots, X_n\rangle$*

Example 6 (Random Walk Martingale).

Example 7 (Multiplicative Martingale).

Example 8 (Moment Generating Function Martingale).

Example 9 (Smoothing Martingale).

Example 10 (Averaging Martingale).

Example 11 (Radon-Nikodym Derivative Martingale).

Theorem 187 (Some basic transformations).

1. *If X_n and Y_n are submartingales wrt a filtration $\mathcal{F}_n$, then so are $X_n + Y_n$ and $\max(X_n, Y_n)$.*
2. *Suppose X_n is a martingale wrt a filtration $\mathcal{F}_n$, $f\colon \mathbb{R} \to \mathbb{R}$ is convex, and $Y_n := f(X_n)$ is integrable for each $n \in \mathbb{N}$. Then Y_n is a submartingale wrt $\mathcal{F}_n$.*
3. *Suppose X_n is a submartingale wrt a filtration $\mathcal{F}_n$, $f\colon \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing, and $Y_n := f(X_n)$ is integrable for each $n \in \mathbb{N}$. Then Y_n is a submartingale wrt $\mathcal{F}_n$.*
4. *If $X_{n,1}, \ldots, X_{n,k}$ are k submartingales wrt a common filtration $\mathcal{F}_n$ and $w_1, \ldots, w_k$ are nonnegative weights, then the process $Y_n := \sum_{j=1}^{k} w_j X_{n,j}$ is also a submartingale wrt $\mathcal{F}_n$.*

Example 12 (Gambler's strategy).

21 Stopping times and the optional sampling theorem

Definition 94. *An extended random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ is called a* **stopping time** *with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if*

$$\{\tau = n\} \in \mathcal{F}_n \text{ for all } n \in \mathbb{N}.$$

Definition 95 (The information at τ). *Let τ be a stopping time wrt a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Then $\mathcal{F}_\tau$ denotes the, so called,* **pre-τ σ-field**, *which is defined as the set of all subsets A of Ω such that*

$$A \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n \in \mathbb{N} \text{ and}$$
$$A \cap \{\tau = \infty\} \in \mathcal{F}.$$

Theorem 188 (Basic properties of the stopped σ-field). *Suppose τ and σ are stopping times wrt a a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Then,*

1. *$\mathcal{F}_\tau$ is a sub σ-field of $\mathcal{F}$ and*
$$\begin{aligned} A \in \mathcal{F}_\tau &\iff A \in \mathcal{F} \text{ and } A \cap \{\tau = n\} \in \mathcal{F}_n, \forall n \in \mathbb{N} \\ &\iff A \in \mathcal{F} \text{ and } A \cap \{\tau \leq n\} \in \mathcal{F}_n, \forall n \in \mathbb{N}; \end{aligned}$$
2. *$\{\sigma \leq \tau\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$;*
3. *If $\{\sigma \leq \tau\} = \Omega$ then $\mathcal{F}_\sigma \subset \mathcal{F}_\tau$;*
4. *If $\tau < \infty$ and $(X_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ then X_τ is $\mathcal{F}_\tau$-measurable;*
5. *If $(X_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ then $Y := \inf\{X_{\tau\wedge n} \colon n \in \mathbb{N}\}$ is $\mathcal{F}_\tau$-measurable.*

Theorem 189 (Finite optional sampling (FOST)). *Let $(X_1, \ldots, X_n)$ be a submartingale wrt a filtration $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$. Let σ and τ be stopping times wrt $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$ such that $\sigma \leq \tau \leq n$. Then*

$$(X_\sigma, X_\tau) \text{ is a submartingale wrt the filtration } (\mathcal{F}_\sigma, \mathcal{F}_\tau).$$

Theorem 190 (Kolmogorov's inequality for submartingales). *Let $(X_1, \ldots, X_n)$ be a submartingale wrt the filtration $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$ and set $M_n := \max(X_1, \ldots, X_n)$. For each $c > 0$ one has*

$$P(M_n \geq c) \leq \frac{E(X_n I_{\{M_n \geq c\}})}{c} \leq \frac{E(X_n^+)}{c}. \tag{92}$$

Theorem 191 (Azuma's inequality). *Let $(X_1, \ldots, X_n)$ be a martingale wrt the filtration $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$ and set $M_n := \max(X_1, \ldots, X_n)$. In addition, for each $k = 1, \ldots, n$, suppose $E(X_k) = 0$ and there exists finite $\alpha_k, \beta_k > 0$ such that*

$$-\alpha_k \leq X_k - X_{k-1} \leq \beta_k.$$

Then for each $c > 0$

$$P(M_n \geq c) \leq \inf_{a>0}\Big(e^{-ac}\prod_{k=1}^{n}\frac{\beta_k e^{-\alpha_k a} + \alpha_k e^{\beta_k a}}{\alpha_k + \beta_k}\Big) \leq \exp\Big(-\frac{c^2}{2\tau^2}\Big)$$

where $\tau^2 := \frac{1}{4}\sum_{k=1}^n (\alpha_k + \beta_k)^2$

Theorem 192 (Hoeffding's inequality). *Let $D_1, \ldots, D_n$ be independent bounded random variables such that $D_k \in [a_k, b_k]$ for finite $a_k \leq b_k$, $k = 1, \ldots, n$. Let $S_n := D_1 + \cdots + D_n$. Then for any $c > 0$*

$$P\big(S_n - ES_n \geq c\big) \leq \exp\Big(-\frac{c^2}{2\tau^2}\Big)$$

where $\tau^2 := \frac{1}{4}\sum_{k=1}^n (b_k - a_k)^2$.

Theorem 193 (McDiarmid's inequality). *Let $D_1, \ldots, D_n$ be independent random variables taking values in ranges $R_1, \ldots, R_n$. Let $F\colon R_1 \times \cdots \times R_n \to \mathbb{R}$ have the property that for all $k = 1, \ldots, n$ there exists a finite constant $c_k > 0$ such that*

$$\begin{aligned}\big|F(x_1, \ldots, x_{k-1}, a, &x_{k+1}, \ldots, x_n)\\ &- F(x_1, \ldots, x_{k-1}, b, x_{k+1}, \ldots, x_n)\big| \leq c_k\end{aligned}$$

for all $a, b \in R_k$ and $x_j \in R_j$. Then for any $c > 0$

$$P\big(F(D_1, \ldots, D_n) - E(F(D_1, \ldots, D_n)) \geq c\big) \leq \exp\Big(-\frac{c^2}{2\tau^2}\Big)$$

where $\tau^2 := \sum_{k=1}^n c_k^2$.

Theorem 194 (Doob's upcrossing). *If $(X_1, \ldots, X_n)$ is a non-negative submartingale wrt the filtration $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$, then for every $c > 0$ the number U of upcrossings of $[0, c]$ satisfies*

$$E(U) \leq \frac{E(X_n) - E(X_1)}{c} \tag{93}$$

Corollary 26. *If $(X_1, \ldots, X_n)$ is a submartingale wrt the filtration $(\mathcal{F}_1, \ldots, \mathcal{F}_n)$, then for every $b > a$ the number U of upcrossings of $[a, b]$ satisfies*

$$E(U) \leq \frac{E(X_n - a)^+ - E(X_1 - a)^+}{b - a} \leq \frac{E(X_n^+) + a^-}{b - a}$$

22 Martingale Limit Theorems

Section Assumption. *For the remainder of this section let $(X_n)_{n\in\mathbb{N}}$ be a submartingale wrt the filtration $\mathcal{F}_n$ and let $\mathcal{F}_\infty := \sigma\langle \mathcal{F}_1, \mathcal{F}_2, \ldots\rangle$.*

Theorem 195 (Almost sure convergence (ASCT)). *If $\sup_n E(X_n^+) < \infty$ then there exists an $\mathcal{F}_\infty$-measurable and integrable random variable X_∞ such that*

$$X_n \xrightarrow{ae} X_\infty.$$

Theorem 196 (Martingale smoothing). *Let X be an $\mathcal{F}$-measurable integrable random variable. Then the collection of random variables $(E^{\mathcal{F}_n}X)_{n\in\mathcal{F}_n}$ is UI and*

$$E^{\mathcal{F}_n}X \longrightarrow E^{\mathcal{F}_\infty}X$$

a.e. and in L_1 as $n \to \infty$.

Definition 96 (A closer). *A pair $(X_\bullet, \mathcal{F}_\bullet)$ consisting of a random variable $X_\bullet$ and a sub-σ-field $\mathcal{F}_\bullet$ of $\mathcal{F}$ is said to* **close the sub-martingale** $(X_n)_{n\in\mathbb{N}}$ **on the right** *if*

$$X_1, X_2, \ldots, X_n, \ldots, X_\bullet$$

is a sub-martingale wrt the filtration

$$\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \ldots, \mathcal{F}_\bullet.$$

The pair $(X_\circ, \mathcal{F}_\circ)$ is said to be **the nearest closer of** $(X_n)_{n\in\mathbb{N}}$ **on the right** *if*

$$X_1, X_2, \ldots, X_n, \ldots, X_\circ, X_\bullet$$

is a sub-martingale wrt the filtration

$$\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \ldots, \mathcal{F}_\circ, \mathcal{F}_\bullet$$

for every closer $(X_\bullet, \mathcal{F}_\bullet)$

Theorem 197 (A closer of $(X_n)_{n\in\mathbb{N}}$). *If there exists a closer of $(X_n)_{n\in\mathbb{N}}$ then there exists an $\mathcal{F}_\infty$-measurable and integrable random variable X_∞ such that*

$$X_n \xrightarrow{ae} X_\infty.$$

and X_∞ is the nearest closer of $(X_n)_{n\in\mathbb{N}}$.

Theorem 198 (subM L_p convergence theorem). *If $|X_n|^p$ are UI where $1 \leq p < \infty$ then there exists an $\mathcal{F}_\infty$-measurable random variable X_∞ such that $X_\infty \in L_p$*

$$X_n \xrightarrow{ae} X_\infty \text{ and } X_n \xrightarrow{L_p} X_\infty$$

and X_∞ is the nearest closer of $(X_n)_{n\in\mathbb{N}}$.

Theorem 199 (Equivalence of some convergence criterion).

- *There exists a closer of $(X_n)_{n\in\mathbb{N}} \iff X_n^+$ are UI*

- *If the X_n's are non-negative and $p > 1$ then*

$$|X_n|^p \text{ are UI} \iff \sup_n E(|X_n|^p) < \infty \iff E(\sup_n |X_n|^p) < \infty$$

Theorem 200 (**Application to likelihood ratios**). *Let Q be another probability measure on $(\Omega, \mathcal{F})$. For $n \in \mathbb{N} \cup \{\infty\}$ define*

$$Q_n := Q\big|_{\mathcal{F}_n} \text{ and } P_n := P\big|_{\mathcal{F}_n}.$$

Consider the Lebesque decomposition of Q_n with respect to P_n:

$$Q_n(\bullet) = Q_n^a(\bullet) + Q_n^s(\bullet) = \int_\bullet \rho_n dP_n + Q_n(\bullet \cap N_n)$$

where $\rho_n = \frac{dQ_n^a}{dP_n}$ and N_n is P_n-null. Then $(\rho_n)_{n\in\mathbb{N}}$ is a non-negative super-martinagle and

$$\rho_n \xrightarrow{ae} \rho_\infty$$

Notice that when $\mathcal{F}_n = \sigma\langle X_1, \ldots, X_n\rangle$ for *any* sequence of random variables $X_1, X_2, \ldots$ on $(\Omega, \mathcal{F}, P)$, not just (sub)martingales, then ρ_n has the form

$$\rho_n =_{a.e.} I_{\{p_n(X_1,\ldots,X_n)>0\}} \frac{q_n(X_1, \ldots, X_n)}{p_n(X_1, \ldots, X_n)}$$

where q_n and p_n are densities of $P_n(X_1, \ldots, X_n)^{-1}$ and $Q_n(X_1, \ldots, X_n)^{-1}$ with respect to some measure μ_n, respectively. Also, note that there always exists some such measure μ_n since one can take $\mu_n = Q_n(X_1, \ldots, X_n)^{-1} + P_n(X_1, \ldots, X_n)^{-1}$.

Exercise 59. *Let $X_1, X_2, \ldots$ be random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let Q be another probabilty measure on $(\Omega, \mathcal{F})$. Let $\mathcal{F}_n := \sigma\langle X_1, \ldots, X_n\rangle$, $\mathcal{F}_\infty := \sigma\langle X_n : n \in \mathbb{N}\rangle$ and*

$$P_n = P\big|_{\mathcal{F}_n} \text{ and } Q_n = Q\big|_{\mathcal{F}_n}$$

for all $n \in \mathbb{N} \cup \{\infty\}$. Let $Q_n = Q_n^a + Q_n^s$ be the Lebesque decomposition of Q_n with respect to P_n and

$$\rho_n := \frac{dQ_n^a}{dP_n} \text{ for all } n \in \mathbb{N} \cup \{\infty\}.$$

- *Show that the process $(\sqrt{\rho_n})_{n\in\mathbb{N}}$ is UI, is in $L_2(P)$ and is a super-martingale.*
- *Show that $E(\sqrt{\rho_n}) \downarrow E(\sqrt{\rho_\infty})$ as $n \to \infty$.*
- *Show that $Q_\infty \perp P_\infty \iff \lim_n E(\sqrt{\rho_n}) = 0$.*
- *Show that the following statements are equivalent*
  1. *$Q_\infty \ll P_\infty$*
  2. *$Q_n \ll P_n$ for all $n \in \mathbb{N}$ and the ρ_n's are UI*
  3. *$Q_n \ll P_n$ for all $n \in \mathbb{N}$ and the $\sqrt{\rho_n}$'s converge in L_2*
  4. *$Q_n \ll P_n$ for all $n \in \mathbb{N}$ and the $\sqrt{\rho_n}$'s are Cauchy in L_2*
  5. *$Q_n \ll P_n$ for all $n \in \mathbb{N}$ and the $\lim_{n,m} E(\sqrt{\rho_m}\sqrt{\rho_n}) = 1$.*

  Remark: The condition $\lim_{n,m} E(\sqrt{\rho_m}\sqrt{\rho_n}) = 1$ is related to a Cauchy criterion for Hellinger distance.

23 Backward sub-martingales

Definition 97 (**Backward sub-martingales**). *A submartingale indexed by the negative integers is called a backward sub-martingale. In particular, $(X_{-n})_{n\in\mathbb{N}}$ is said to be* **a backward sub-martingale with respect to filtration** $(\mathcal{F}_{-n})_{n\in\mathbb{N}}$ *if for each $n \in \mathbb{N}$*

- *$\mathcal{F}_{-n} \subset F_{-n+1}$*
- *X_{-n} is $\mathcal{F}_{-n}$-measurable and integrable*
- *$E^{\mathcal{F}_{-n}}(X_{-n+1}) \geq_{a.e.} X_{-n}$*

Theorem 201 (**Backward almost sure convergence**). *If $(X_{-n})_{n\in\mathbb{N}}$ is a backward sub-martingale with respect to filtration $(\mathcal{F}_{-n})_{n\in\mathbb{N}}$ then there exists an extended random variable $X_{-\infty}$ such that*

$$X_n \xrightarrow{ae} X_\infty.$$

as $n \to \infty$ where $X_{-\infty} \in Q^+$ and is measurable with respect to

$$\mathcal{F}_{-\infty} := \bigcap_{n\in\mathbb{N}} \mathcal{F}_{-n}.$$

Theorem 202 (**Backward closer**). *If $(X_{-n})_{n\in\mathbb{N}}$ is a backward sub-martingale with respect to filtration $(\mathcal{F}_{-n})_{n\in\mathbb{N}}$ which has a left closer then there exists $X_{-\infty} \in L_1(P, \mathcal{F}_{-\infty})$ such that*

$$X_{-n} \xrightarrow{ae} X_{-\infty} \text{ and } X_{-n} \xrightarrow{L_1} X_{-\infty}$$

as $n \to \infty$ where $X_{-\infty}$ is the nearest left closer of $(X_{-n})_{n\in\mathbb{N}}$.

24 Continuous time martingales

Part VI
Markov Chains

25 Basic Theory

26 Brownian motion

27 Skorokhod's embedding

Part VII

Probability Inequalities

28 Maximal inequalities

29 Concentration of measure

Part VIII
Stochastic processes

30 Constructing probability measures on infinite product spaces

30.1 Transition probabilities

30.2 Tulcea's theorem

Requires a countable set of transition probabilities.

30.3 Product probability theorem

Requires independent marginals.

30.4 Kolmogorov's extension theorem

Requires consistency of all finite dimensional marginals.

31 Gaussian random fields

31.1 Metric embeddings, Hilbert spaces and Schoenberg's results

31.2 Dirichlet forms, Green's function, resistance metric, Markov chain characterizations

31.3 Reproducing kernels, and a set of isometric Hilbert spaces

32 Karhunen-Loève

33 Stationary processes

34 White noise

35 SDE and Integration with respect to white noise

Part IX
Empirical Process Theory

36 Dudley's chaining argument

37 Empirical process theory