Probability Theory Notes

Zhang Ruiyang

Contents

Preface

This is the notes taken while taking MATH0069 Probability taught by Prof Nadia Sidorova at UCL in Term 2, 2023.

In the first chapter, we will cover the foundation of probability theory, i.e. define a probability space, a random variable etc. It would be mostly tedious technical work but they are necessary preliminary work needed for the following content. In Chapter 2, we will study the concept of independence, which is the key idea that makes probability theory not mere measure theory. We will also introduce the concept of tail events and the interesting result of Kolmogorov 0-1 law. In Chapter 3, we will cover two key and central results in probability - the (strong) law of large numbers and the central limit theorem. In the final chapter, we will dive into martingales, which is a particularly nice class of stochastic processes with interesting properties.

Familiarities with measure theory are compulsory, and we will use results from measure theory directly every now and then. Some essential definitions and results are contained in the Appendix, such as the monotone convergence theorem and dominated convergence theorem.

The course (and therefore this notes) is based on Williams' *Probability with Martingales.*

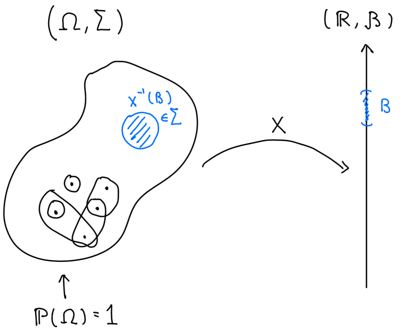
Chapter 1

Rigorous Introduction to Basic Probability

1.1 Probability Space and Random Variables

Definition 1.1. *$(\Omega, \Sigma, \mathbb{P})$, where (Ω, Σ) is a measure space and $\mathbb{P}$ is a probability measure (i.e. measure with $\mathbb{P}(\Omega) = 1$) on (Ω, Σ), is a **probability space**. Furthermore, we call Ω as the **sample space**, and Σ as the **event space**.*

Definition 1.2. *A measurable function $X : \Omega \to \mathbb{R}$, equipped with Σ and the completed Borel σ-algebra $\mathcal{B}$, is called a **random variable**.*



This is the formal definition of a random variable. Let us see some familiar and unfamiliar examples to solidify our understanding of this concept.

Example. Coin Toss.

(a) $\Omega = \{0, 1\}$, $\Sigma = 2^\Omega$, and $\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = 1/2$. The desired random variable X is defined by $0 \mapsto 0$ and $1 \mapsto 1$.

(b) $\Omega = [0, 1]$, $\Sigma = \mathcal{B}[0, 1]$, $\mathbb{P} = Leb$. The random variable is defined by

$$X(\omega) = \begin{cases} 0 & \omega \in [0, \frac{1}{2}) \\ 1 & \omega \in [\frac{1}{2}, 1]. \end{cases}$$

The two characterisations of the random variable are in fact identical, in the sense that they yield the same law (and thus distribution) of the random variable. The exact meanings of these terms will be clarified in the following section. At this stage, we will just need to understand that they are alternative, but equivalent, ways to define a probability space and a random variable that mean the same thing. In practice, we will use the most convenient characterisation, depending on our purposes.

Example. Roll a dice, spell the number, and take the number of letters.

(a) $\Omega = \{1,2,3,4,5,6\}$, $\Sigma = 2^\Omega$, and $\mathbb{P}(\{1\}) = \cdots = \mathbb{P}(\{6\}) = 1/6$, extended by additivity. The desired random variable X is defined by

$$X : \begin{cases} 1 \mapsto 3 \\ 2 \mapsto 3 \\ 3 \mapsto 5 \\ 4 \mapsto 4 \\ 5 \mapsto 4 \\ 6 \mapsto 3. \end{cases}$$

(b) $\Omega = \{0,1,2,3,4,5,6\}$, $\Sigma = 2^\Omega$, and $\mathbb{P}(\{0\}) = 0, \mathbb{P}(\{1\}) = \cdots = \mathbb{P}(\{6\}) = 1/6$, extended by additivity. The random variable is defined the same as above, and it maps 0 to any arbitrary thing. It would not matter as 0 occurring has probability 0.

(c) $\Omega = \{3,4,5\}$, $\Sigma = 2^\Omega$, and $\mathbb{P}(\{3\}) = 1/2, \mathbb{P}(\{4\}) = 1/3, \mathbb{P}(\{5\}) = 1/6$, extended by additivity. The random variable is defined to be the identity map on Ω.

As above, the goal of this example is to illustrate the point that different characteristics could produce the same random variable.

Example. Toss 2 coins, and take the sum of the outcomes.

$\Omega = \{0,1\} \times \{0,1\}$, $\Sigma = 2^\Omega$, $\mathbb{P}(\{\text{each point}\}) = 1/4$, extended by additivity. The random variable is defined by

$$X : \begin{cases} (0,0) \mapsto 0 \\ (0,1) \mapsto 1 \\ (1,0) \mapsto 1 \\ (1,1) \mapsto 2. \end{cases}$$

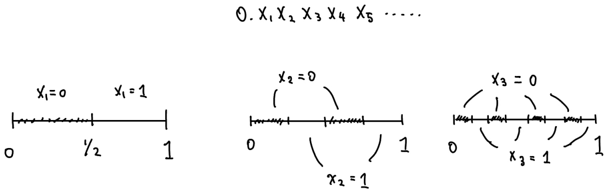
The above example is not too hard. But what if we toss, not two but, an infinite number of coins consecutively? What would things look like? A natural way of extending to this case is to consider Ω to be the infinite Cartesian product of $\{0,1\}$, and define other things accordingly. There are a lot of technical details that we need to take care of, for example we could not use the simple power set as our Σ as Ω is no longer finite. Those technical difficulties could be overcome, but we do not do them here. Instead, we have the following characterisation of this random variable that involves a neat little trick.

Example. Toss a coin consecutively, so a possible outcome would be HTHHTHTHTHT ...

The idea behind these characteristics is the following. We will denote each possible outcome as $0.x_1x_2x_3x_4\ldots$ where $x_i \in \{0,1\}$. This is a binary number in $[0,1]$. Furthermore, each outcome corresponds to a point in this closed interval. The correspondence is as the following: We will cut the unit interval into two halves, pick left if the first digit after the decimal point is 0, and pick right if it is 1. Then, we will cut the interval we pick into halves, pick left if the second digit is 0 and pick right if it is 1. And it goes on like that.

One potential issue is that two outcomes might correspond to the same point. For example, 0.01 and 0.0011111... will correspond to the same point on the unit interval. This is not an actual issue. Because the set of such troublesome points (e.g. $1/2, 1/4, 3/4$) forms a countable set, so it has Lebesgue measure 0 and therefore would not affect our set-up.

Therefore, we have $\Omega = [0,1]$, $\Sigma = \mathcal{B}[0,1]$, $\mathbb{P} = Leb$. A potential event, say the event of the first toss being head, corresponds to the right half of the unit interval in the Ω by the way we construct it. The event of the second toss being tail then corresponds to the first and third (from the left) quarters of the unit interval.



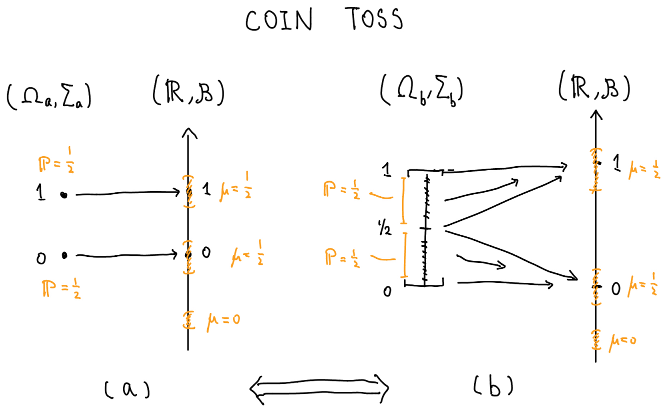
The random variable will be to map a point in $[0,1]$ to the reverse-engineered string of binary numbers corresponding to heads and tails.

1.2 Law and Distribution of Random Variables

Definition 1.3. *$\mu(B) := \mathbb{P}(X^{-1}(B))$ for all $B \in \mathcal{B}$ is the* ***law*** *of the random variable X. $F(X) := \mu((-\infty, x])$ for $x \in \mathbb{R}$ is the* ***distribution function*** *of random variable X.*

The idea behind the second form is that $(-\infty, x]$ with all possible x is a generating set for the Borel set $\mathcal{B}$ too, and that form of the function is much easier to deal with than the first form.

Essentially, if two objects have the same law (and thus the same distribution function), then they are the same random variable. Using this, we can confirm our previous remark on the equivalence of the various characteristics of random variables for the examples. The following diagram illustrates roughly why the two characterisations of the coin toss random variable are equivalent.



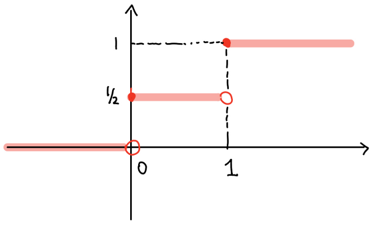
We will sometimes use the **Dirac measure** at x, which we denote it as δ_x, and it is defined as

$$\delta_x(B) = \begin{cases} 1 & x \in B \\ 0 & x \notin B. \end{cases}$$

So, the law μ of the above coin toss random variable can also be written as

$$\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

We can also draw the distribution function for the above coin toss example.



This distribution function has some properties, such as it is increasing and it is between 0 and 1. In fact, there are several universal properties of a distribution function.

Theorem 1.4 (Properties of Distribution Function)**.** *We have, for any distribution function F,*

1. *F is increasing.*
2. *$\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.*
3. *F is right-continuous.*

Proof. (1) For $s < t$, we have

$$F(s) = \mu((-\infty, s]) \leq \mu((-\infty, t]) = F(t)$$

where the inequality is due to the monotonicity of measure, and the two equalities are simply definitions.

(2) Take $\{t_n\}$ with $t_n \nearrow \infty$[1]. We have

$$\lim_{n\to\infty} F(t_n) = \lim_{n\to\infty} \mu((-\infty, t_n]) = \mu(\cup_{i=1}^{\infty}(-\infty, t_n]) = \mu(\mathbb{R}) = 1$$

where the second equality uses the continuity of measure, and the last equality uses the fact that μ is a probability measure. Similarly, take $\{t_n\}$ with $t_n \searrow -\infty$[2]. We have

$$\lim_{n\to\infty} F(t_n) = \lim_{n\to\infty} \mu((-\infty, t_n]) = \mu(\cap_{i=1}^{\infty}(-\infty, t_n]) = \mu(\emptyset) = 0$$

where the second equality uses the continuity of measure.

(3) Let $t \in \mathbb{R}$, and take $\{t_n\}$ with $t_n \to t^+$[3]. We have

$$\lim_{n\to\infty} F(t_n) = \lim_{n\to\infty} \mu((-\infty, t_n]) = \mu(\cap_{i=1}^{\infty}(-\infty, t_n]) = \mu((-\infty, t]) = F(t).$$

□

[1] $t_k \leq t_{k+1}$ for all k, and $t_k \to \infty$ as $k \to \infty$.

[2] $t_k \geq t_{k+1}$ for all k, and $t_k \to -\infty$ as $k \to \infty$.

[3] $t_k \geq t$ for all k and $t_k \to t$ as $k \to \infty$.

Remark. The reason why F is right-continuous is based on the way it is defined, i.e. $F(X) := \mu((-\infty, x])$. If we define it as $F(X) := \mu((-\infty, x))$, which is perfectly fine, we would have F being left-continuous. The version of F with right-continuity was the version chosen by Kolmogorov for whatever reason, therefore it is the version that everyone uses.

We know that given a random variable, we can find a distribution function for it. It turns out that the reverse is true, i.e. given a distribution function we can find a random variable for it. The construction that takes a distribution function and produces a random variable is called **Skorokhod Construction**.

Theorem 1.5 (Skorokhod Construction)**.** *Let $F : \mathbb{R} \to \mathbb{R}$ be a function satisfying the properties of a distribution function as shown in Theorem 1.4. Then, there exists a random variable X such that F is its distribution function.*

Remark. This result allows us to work with just the distribution functions and to not worry about the probability space which could be troublesome to determine sometimes.

Proof. Consider $([0, 1], \mathcal{B}[0, 1], Leb)$ to be the probability space that this random variable is in.

We will first prove an easy version of this result with additional conditions on F. This illustrates the essence of the proof. Next, we will remove those conditions and prove the original statement.

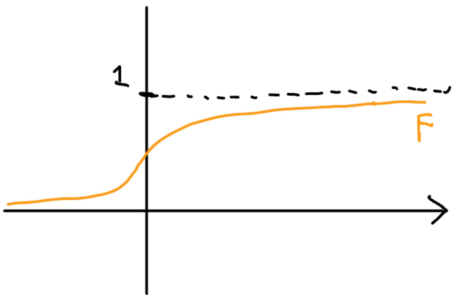
Suppose F is, in addition, strictly increasing and continuous. This means that the inverse of this function is well-defined. The desired random variable is then simply

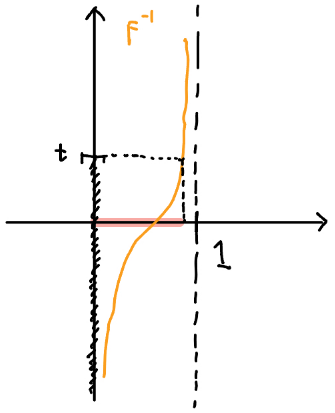
$$X(\omega) = \begin{cases} F^{-1}(\omega), & \omega \in (0, 1) \\ \text{anything}, & \omega \in \{0, 1\}. \end{cases}$$

We let F_X denotes the distribution function of X. We have

$$\begin{aligned} F_X(t) &= \mu_X((-\infty, t]) \\ &= P(X^{-1}((-\infty, t])) \\ &= Leb(\{\omega \in [0, 1] \mid X(\omega) \in (-\infty, t]\}) := Leb(\{X \leq t\}) \\ &= F(t) \end{aligned}$$

as desired.





Now we will prove this theorem without the extra condition.

There are two things that will be problematic if we want to use the same approach. The first thing is a constant piece in the distribution, and the second is a jump in the distribution. These two things avoid us from taking the straightforward inverse. The good news is that we can work around them.

We define the function G that works as a pseudo-inverse of F. We have

$$G(\omega) = \inf\{u \mid F(u) > \omega\}$$

What this definition does is that it converts a constant path into a jump, and a jump to a constant path taking the position after the jump.

We define our random variable $X(\omega) = G(\omega)$, and we would like to check that it has the desired distribution function. This is the same as checking

$$[0, F(t)) \subset \{\omega \in [0, 1] \mid X(\omega) \in (-\infty, t]\} \subset [0, F(t)]$$

which is equivalent to the desired result after taking the Lebesgue measure to everything (notice that the middle quantity will be bounded by $F(t)$, thus is equal to $F(t)$).

For the first relationship, for $\omega \in [0, F(t))$, we have

$$X(\omega) = \inf\{u \mid F(u) > \omega\} \leq t$$

so $X(\omega) \in (-\infty, t]$. For the second relationship, if we have ω with $X(\omega) \leq t$, i.e. $\inf\{u \mid F(u) > \omega\} \leq t$, then we can apply F to both sides and get

$$\inf\{F(u) \mid F(u) > \omega\} \leq F(t) \implies \omega \leq F(t),$$

so $\omega \in\subset [0, F(t)]$. Done. □

1.3 Examples of Random Variables and Their Distributions

Definition 1.6. *A random variable with distribution function F is called a* ***continuous random variable*** *if F can be written as*

$$F(t) = \int_{-\infty}^{t} f(s)ds$$

for some Lebesgue integrable function f and f is called the ***density****. We would also say F admits a density f to mean the same thing.*

Naturally, if F admits a density f, then we would have $f = F'$, i.e. f is the derivative of F. Additionally, each f would have the property

$$\int_{-\infty}^{\infty} f(s)ds = 1$$

as $\lim_{t\to\infty} F(t) = 1$.

Definition 1.7. *A random variable X is called a* ***discrete random variable*** *if there exists finite or countably many $a_1, a_2, \ldots \in \mathbb{R}$ and $p_1, p_2 \ldots \in [0,1]$ with $\sum p_i = 1$ such that we have*

$$\mathbb{P}(X = a_i) = p_i \qquad\qquad \forall i = 1, 2, \ldots.$$

In this case, the distribution function F of X is defined to be

$$F(x) = \mathbb{P}(X \leq x) = \sum_{i : a_i \leq x} \mathbb{P}(X = a_i) = \sum_{i : a_i \leq x} p_i.$$

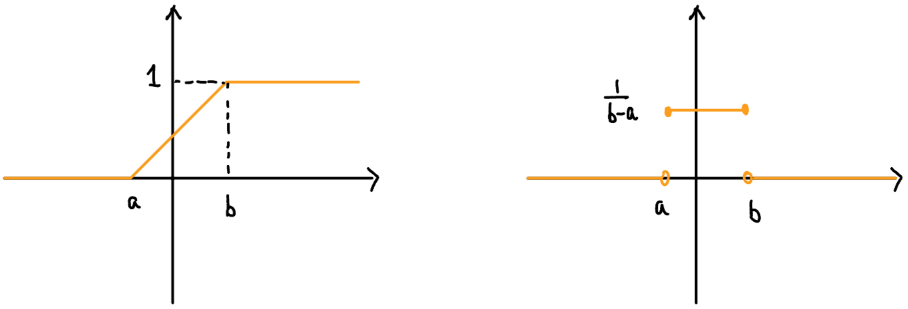
Now, we will provide some basic examples of random variables, which should be familiar.

Example. Uniform random variable on $[a, b]$, denoted by $Unif[a, b]$. It has distribution function F defined by

$$F(x) = \begin{cases} 1 & x > b \\ \frac{x-a}{b-a} & x \in [a, b] \\ 0 & x < a \end{cases}$$

and it admits the density f defined by

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b]. \end{cases}$$



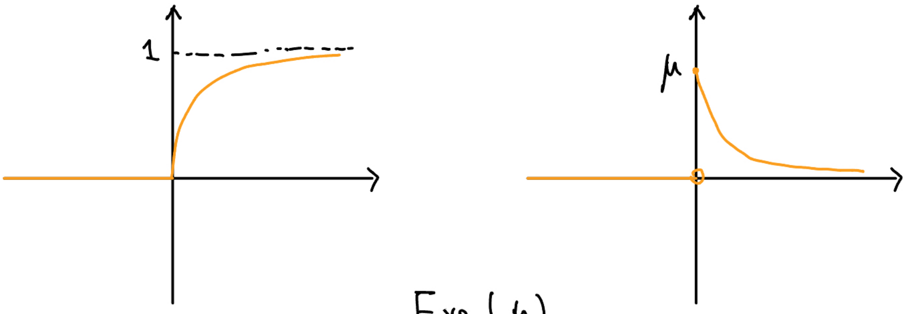
Unif [a,b]

Example. Exponential random variable with parameter μ, denoted by $Exp(\mu)$. It has distribution function F defined by

$$F(x) = \begin{cases} 1 - e^{-x/\mu} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

and it admits the density f defined by

$$f(x) = \begin{cases} \mu e^{-x/\mu} & x > 0 \\ 0 & x \leq 0. \end{cases}$$



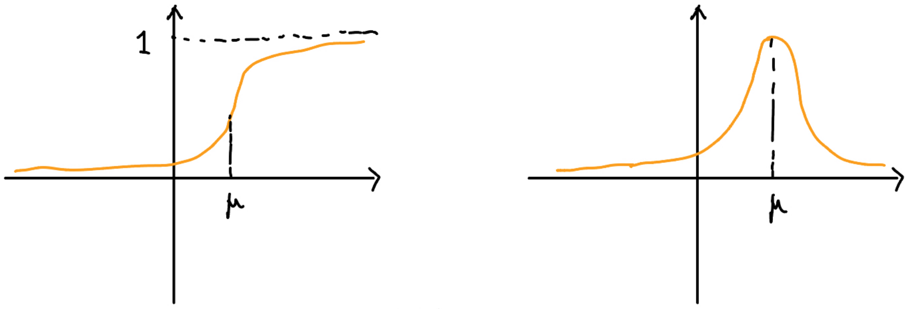
Exp (μ)

Example. Normal random variable with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$. It has distribution function F defined by

$$F(x) = \int_{-\infty}^{x} f(s)ds$$

and it admits the density f defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$



N(μ, σ²)

The three examples above are continuous random variables. The two examples below are discrete random variables.

Example. Bernoulli random variable X with success probability p, denoted by $Ber(p)$. It takes 0 and 1, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

Example. Poisson random variable X with mean μ, denoted by $Poi(\mu)$. It takes natural numbers, with $\mathbb{P}(X = k) = e^{-\mu}\mu^k/k!$ for all $k = 0, 1, 2, \dots$.

Of course, we could have random variables that are neither discrete nor continuous. A trivial way to obtain such a random variable is to be a mixture of a discrete random variable and a continuous random variable, for example this random variable X will be like a $Unif[0, 1]$ with probability 1/2 and be like a $Ber(0.1)$ with probability 1/2. However, we can have something constructed less trivially. For example, a uniform distribution on the Cantor set.

1.4 Expectation

Definition 1.8. *Let X be Lebesgue integrable. Then,*

$$\mathbb{E}(X) = \int_{\Omega} X d\mathbb{P}$$

is called the ***expectation*** *of X. If $X \geq 0$ and is not Lebesgue integrable, then $\mathbb{E}(X) = \infty$. Similarly, if $X \leq 0$ and is not Lebesgue integrable, then $\mathbb{E}(X) = -\infty$.*

Theorem 1.9. *Let $h : \mathbb{R} \to \mathbb{R}$ be integrable with regards to μ. Then, we have*

$$\mathbb{E}[h(X)] = \int_{\Omega} h(X) d\mathbb{P} = \int_{\mathbb{R}} h d\mu.$$

Proof. The proof of this is very similar to the proof of Lebesgue integrals. We will do the first part and the rest will be almost identical.

Consider $h = 1_B$ for some $B \in \mathcal{B}$. Then, we have

$$\int_{\Omega} h(X) d\mathbb{P} = \int_{\Omega} 1_B(X) d\mathbb{P} = \mathbb{P}(X \in B)$$

and

$$\int_{\mathbb{R}} h d\mu = \int_{\mathbb{R}} 1_B d\mu = \mu(B),$$

which really are equal by the definition of μ. □

This result also allows us to recover the familiar expectation formula for discrete and random variables. For a discrete random variable X that takes a_i with probability p_i, we have

$$\mathbb{E}[h(X)] = \int h d\mu = \int h(\sum p_i \delta_{a_i}) = \sum p_i h(a_i).$$

For a continuous random variable X with density f, we have

$$\mathbb{E}[h(X)] = \int h d\mu = \int h(x) f(x) dx.$$

A simple yet powerful result is the Markov inequality.

Theorem 1.10 (Markov Inequality)**.** *Let X be a non-negative integrable random variable. Then, for any $c \in \mathbb{R}$, we have*

$$c \cdot \mathbb{P}(X > c) \leq \mathbb{E}[X].$$

Proof. This result is obvious for any $c \leq 0$ since X is non-negative. For $c > 0$, we consider this random variable $Y := c1_{\{X>c\}}$. Notice that $Y \leq X$ almost surely, so we have

$$c \cdot \mathbb{P}(X > c) = \mathbb{E}[Y] \leq \mathbb{E}[X],$$

as desired. □

Corollary 1.11. *Let X be a non-negative integrable random variable with X^n being integrable as well for some integer n. Then, for any $c > 0$, we have*

$$c^n \mathbb{P}(X^n > c^n) \leq \mathbb{E}[X^n].$$

Such inequalities are known as concentration inequalities, and they allow us to provide bounds for the tails of the distribution. This is frequently used in high-dimensional statistics (and probability).

Definition 1.12. *If X is square integrable, i.e. $\mathbb{E}[X^2] < \infty$, then*

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

*is the **variance** of X.*

Lemma 1.13. $\mathbb{E}[X^2] < \infty \implies \mathbb{E}[|X|] < \infty$.

Proof. Using Cauchy-Schwartz, we have

$$\mathbb{E}[|X|] = \mathbb{E}[|X| \cdot 1] \leq \sqrt{\mathbb{E}[X^2]} \cdot \sqrt{1} = \sqrt{\mathbb{E}[X^2]} < \infty,$$

as desired. □

Chapter 2

Independence and Tail Events

Definition 2.1. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space.*

- *σ-algebras $\Sigma_1, \Sigma_2, \ldots, \Sigma_n \subset \Sigma$ are* ***independent*** *if*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$$

  for any $A_1 \in \Sigma_1, \ldots, A_n \in \Sigma_n$. Additionally, σ-algebras $\Sigma_1, \Sigma_2, \ldots \subset \Sigma$ are ***independent*** *if for any m, $\Sigma_1, \Sigma_2, \ldots, \Sigma_m$ are independent.*
- *Events $E_1, E_2, \ldots \in \Sigma$ are* ***independent*** *if their generated σ-algebras*[1] *$\sigma(E_1), \sigma(E_2), \ldots$ are independent.*
- *Random variables $X_1, X_2, \ldots$ on $(\Omega, \Sigma, \mathbb{P})$ are* ***independent*** *if their generated σ-algebras*[2] *$\sigma(X_1), \sigma(X_2), \ldots$ are independent.*

For two objects A, B that are independent, we will denote it by $A \perp B$.

2.1 π-Systems

This definition is obviously too cumbersome to check in reality, especially when we would like to check for the independence of events and random variables.

For events $A_1, \ldots, A_n$, to check for independence, we just need to check that

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m}) = \mathbb{P}(A_{k_1})\mathbb{P}(A_{k_2}) \cdots \mathbb{P}(A_{k_m})$$

for all possible $\{k_1, k_2, \ldots, k_m\} \subset \{1, 2, \ldots, n\}$. So, for events A, B, C, we just need to check for the triple and all pairs.

This is fine since many possible combinations are equivalent to some of the above. This inspires the following.

Definition 2.2. *$\mathcal{I} \subset \Sigma$ is called a* ***π-system*** *if we have*

1. *$\emptyset \in \mathcal{I}$*
2. *$A, B \in \mathcal{I} \implies A \cap B \in \mathcal{I}$.*

[1] $\sigma(E) = \{E, \Omega \backslash E, \emptyset, \Omega\}$

[2] $\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}\}$

Example. Consider the base set $\mathbb{R}$. $\mathcal{I} = \{(-\infty, x] \mid x \in \mathbb{R}\} \cup \{\emptyset\}$ is a π-system and $\sigma(\mathcal{I}) = \mathcal{B}$. Alternatively, we could have $(-\infty, x)$ instead of $(-\infty, x]$, or (a, b) for $a < b \in \mathbb{R}$. Both these two alternatives will generate $\mathcal{B}$.

Example. Assume $\emptyset$ is included, all closed rectangles in $\mathbb{R}^2$ form a π-system, while all closed discs in $\mathbb{R}^2$ do not.

Example. Consider the probability space $(\Omega, \Sigma, \mathbb{P})$, and let X be a random variable on it. The σ-algebra of X is $\sigma(X) = \{\{X \in B\} \mid B \in \mathcal{B}\}$. If we use the π-system from the above example instead of B, we could get a π-system too. So, we have $\mathcal{I} = \{\{X \leq x\} \mid x \in \mathbb{R}\} \cup \{\emptyset\}$, and $\sigma(\mathcal{I}) = \sigma(X)$.

Example. Consider the probability space $(\Omega, \Sigma, \mathbb{P})$, and let X be a discrete random variable on it with values $a_1, a_2, \ldots$. Similar to the above example, we have $\mathcal{I} = \{\{X = a_1\}, \{X = a_2\}, \ldots\} \cup \{\emptyset\}$ and $\sigma(\mathcal{I}) = \sigma(X)$.

Why should we care about π-systems? Well, given two π-systems $\mathcal{I}$ and $\mathcal{J}$, with $\sigma(\mathcal{I}) = \sigma(X)$ and $\sigma(\mathcal{J}) = \sigma(Y)$ for random variables X and Y, and we have $\mathcal{I} \perp \mathcal{J}$, which implies $\sigma(X) \perp \sigma(Y)$, and therefore $X \perp Y$.

Theorem 2.3. *If π-systems $\mathcal{I}$ and $\mathcal{J}$ are independent, i.e. $\mathbb{P}(I \cap J) = \mathbb{P}(I)\mathbb{P}(J)$ for all $I \in \mathcal{I}$ and $J \in \mathcal{J}$, then their generated σ-algebras $\sigma(\mathcal{I})$ and $\sigma(\mathcal{J})$ are independent.*

The proof requires an auxiliary result from measure theory which we will not prove.

Theorem 2.4. *Let $\mathcal{I}$ be a π-system. Let μ_1, μ_2 be measures on $(\Omega, \sigma(\mathcal{I}))$ such that (1) $\mu_1(\Omega) = \mu_2(\Omega) < \infty$, and (2) $\mu_1 = \mu_2$ on $\mathcal{I}$. Then, we have $\mu_1 = \mu_2$ on $\sigma(\mathcal{I})$.*

Let us prove the desired theorem.

Proof. Fix $J \in \mathcal{J}$. Let $\mu_1(A) = \mathbb{P}(A \cap J)$ where $A \in \sigma(\mathcal{I})$, and $\mu_2(A) = \mathbb{P}(A)\mathbb{P}(J)$ where $A \in \sigma(\mathcal{I})$. Clearly, $\mu_1(\Omega) = \mu_2(\Omega) < \infty$, and $\mu_1 = \mu_2$ on $\mathcal{I}$ by the condition of the theorem. Therefore, $\mu_1 = \mu_2$ on $\sigma(\mathcal{I})$, which means $\mathbb{P}(A \cap J) = \mathbb{P}(A)\mathbb{P}(J)$ for any $A \in \sigma(\mathcal{I})$.

We just need to repeat the process for a fixed $I \in \mathcal{I}$ to obtain the second half of the proof, which we will omit. □

Corollary 2.5. *Two random variables X and Y are independent if and only if the π-systems $\{\{X \leq x\} \mid x \in \mathbb{R}\} \cup \{\emptyset\}$ and $\{\{Y \leq y\} \mid y \in \mathbb{R}\} \cup \{\emptyset\}$ are independent, i.e. we have*

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$$

for all $x, y \in \mathbb{R}$.

Corollary 2.6. *Two discrete random variables X with values $a_1, a_2, \ldots$ and Y with values $b_1, b_2, \ldots$ are independent if and only if*

$$\mathbb{P}(X = a_i, Y = b_j) = \mathbb{P}(X = a_i)\mathbb{P}(Y = b_j).$$

for all possible a_i and b_j.

2.2 Independent Random Variables from Distributions

Recall from the previous chapter that we can construct a random variable given a distribution function (Skorokhod's construction, Theorem 1.5). Now, given a sequence of distribution functions, can we then construct a sequence of random variables such that each follows the respective distribution function and they are independent?

We would not be able to use the same tricks as Skorokhod's constructions as they do not allow us to inject independence into the random variables. Some adjustments are required, and we will start with the case of two distribution functions.

Let us formulate our problem nicely. Consider we have two distribution functions F_1 and F_2, we would like to construct two random variables X_1 and X_2 such that:

1. X_1, X_2 have distribution functions F_1, F_2 respectively.
2. X_1 and X_2 are independent.

First, using Skorokhod construction, we can construct $\tilde{X}_1$ on $(\Omega_1, \Sigma_1, \mathbb{P}_1)$ with distribution function F_1, and $\tilde{X}_2$ on $(\Omega_2, \Sigma_2, \mathbb{P}_2)$ with distribution function F_2. The two probability spaces do not have to be the same.

Next, we will consider the probability space $(\Omega, \Sigma, \mathbb{P}) := (\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ where $\times$ denotes the Cartesian product and $\otimes$ is the product measure. Next, we define two random variables X_1 and X_2 as projections onto their respective spaces and be identical to $\tilde{X}_1$ and $\tilde{X}_2$. This is achieved by defining

$$\begin{aligned} X_1(\omega) &= X_1(\omega_1, \omega_2) = \tilde{X}_1(\omega_1) \\ X_2(\omega) &= X_2(\omega_1, \omega_2) = \tilde{X}_2(\omega_2). \end{aligned}$$

This means we would have, for any $x \in \mathbb{R}$,

$$\mathbb{P}_1 \otimes \mathbb{P}_2(X_1 \leq x) = \mathbb{P}_1 \otimes \mathbb{P}_2(\{X_1 \leq x\} \times \Omega) = \mathbb{P}_1(\{X_1 \leq x\})\mathbb{P}_2(\Omega) = \mathbb{P}_1(\{X_1 \leq x\})$$

where the second last equality follows from the definition of a product measure. We could do the same for the event $X_2 \leq x$. This means the first condition is achieved, i.e. the random variables have their respective distribution functions.

Then, to show X_1 and X_2 are independent, we have, for any $x, y \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}_1 \otimes \mathbb{P}_2(X_1 \leq x, X_2 \leq y) &= \mathbb{P}_1 \otimes \mathbb{P}_2(\{X_1 \leq x\} \times \{X_2 \leq y\}) \\ &= \mathbb{P}_1(\{X_1 \leq x\})\mathbb{P}_2(\{X_2 \leq y\}) \\ &= \mathbb{P}_1 \otimes \mathbb{P}_2(X_1 \leq x)\mathbb{P}_1 \otimes \mathbb{P}_2(X_2 \leq y), \end{aligned}$$

as desired.

Now, we will try to do the same but with a countable sequence of distribution functions $F_1, F_2, \ldots$.

If we attempt to use the same strategy as above, we would require an infinite product of spaces and measures, which is not easy to work with. A trick is thus needed to make sure the construction is still valid. Note that this construction is **not examined**.

Consider we have two distribution functions F_1 and F_2 with their pseudo-inverses as "F_1^{-1}" and "F_2^{-1}" which is the inverse that accounts for jumps and constant pieces. Then, we have two independent random variables U_1 and U_2 following $Unif[0, 1]$. Then, the composed random variable $U_1 \circ$ "F_1^{-1}" and $U_2 \circ$ "F_2^{-1}" will be independent too.

It is easy to construct random variables with the desired distribution functions, as we have the Skorokhod construction. The independence requirement can also be solved now using this uniform random variable trick. The question thus is transformed into constructing a sequence of independent $Unif[0,1]$.

On $([0,1],\mathcal{B},\mathbb{P})$, the uniform random variable $Unif[0,1]$ maps $\omega \in [0,1] \to \omega$. Recall that we can write any number between 0 and 1 using binary digits, and we thus have

$$\omega = 0.\omega_1\omega_2\omega_3\dots$$

where each ω_i is Bernoulli with success rate $1/2$. We can further realise that these ω_is are all independent.

Here comes the key part. We can have, given any $\omega = 0.\omega_1\omega_2\omega_3\dots$

$$\begin{aligned}
&0.\omega_{a_1}\omega_{a_2}\omega_{a_3}\dots\\
&0.\omega_{b_1}\omega_{b_2}\omega_{b_3}\dots\\
&0.\omega_{c_1}\omega_{c_2}\omega_{c_3}\dots\\
&\qquad\qquad\dots
\end{aligned}$$

where $\{a_i\},\{b_i\},\{c_i\},\dots$ are disjoint, and their union is $\mathbb{Z}$. One such construction of such sequence is $\{2^i\},\{3^i\},\{5^i\}\cdots$. Furthermore, each of these values belongs to $[0,1]$ and is independent $Unif[0,1]$. Thus, we have finished the construction.

2.3 Consequences of Independence

Given independence, we can obtain some nice results about the expectation and variance.

Theorem 2.7. *Given independent random variables X and Y, we have*

1. *If $\mathbb{E}[|X|],\mathbb{E}[|Y|] < \infty$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*
2. *If we also have $\mathbb{E}[X^2],\mathbb{E}[Y^2] < \infty$, then $Var[X+Y] = Var[X] + Var[Y]$.*

Proof. The second result is a direct consequence of the first. We have

$$\begin{aligned}
\mathrm{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - [\mathbb{E}(X+Y)]^2\\
&= \mathbb{E}[X^2+2XY+Y^2] - [\mathbb{E}(X)+\mathbb{E}(Y)]^2\\
&= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2\\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2\\
&= \mathrm{Var}[X] + \mathrm{Var}[Y],
\end{aligned}$$

as desired.

The first result can be obtained using the standard routine of measure theory proofs, from indicator functions to simple functions to nonnegative functions to any functions. We will only do the first part.

Consider $X = 1_A$ and $Y = 1_B$ such that X and Y are independent. We have

$$\mathbb{E}[XY] = \int 1_A 1_B d\mathbb{P} = \mathbb{P}(A\cap B),$$

and

$$\mathbb{E}[X]\mathbb{E}[Y] = \mathbb{P}(A)\mathbb{P}(B).$$

Since X and Y are independent, A and B are independent too, so

$$\mathbb{E}[XY] = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \mathbb{E}[X]\mathbb{E}[Y].$$

The rest is routine. □

2.4 Joint Law and Joint Distribution

Definition 2.8. *Let X and Y be random variables defined on the same probability space. Then, the **joint law** on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ is defined as*

$$\mu_{X,Y}(B) = \mathbb{P}((X,Y) \in B).$$

*And, the **joint distribution function** is defined by*

$$F_{X,Y}(x,y) = \mu((-\infty, x] \times (-\infty, y]) = \mathbb{P}(X \le x, Y \le y).$$

Theorem 2.9. *We have*

1. *If X, Y are independent, then $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.*
2. *If X and Y are independent and admit densities f and g, then $\mu_{X,Y}$ admits density with regards to $dxdy$ that is defined by $(x,y) \mapsto f(x)g(y)$, i.e.*

   $$\mu_{X,Y}(B) = \int_B f(x)g(y)\ dxdy$$

   for $B \in \mathcal{B}(\mathbb{R}^2)$.
3. *If X, Y are as in the previous case, then $X + Y$ has density*

   $$f * g(t) = \int_{\mathbb{R}} f(t)g(t-s)ds.$$

Remark. $f * g$ is called the **convolution** of the two functions.

Proof. (1) Notice that $\sigma(X)$ is generated by the π-system $\{\{X \le x\}, x \in \mathbb{R}\}$ and $\sigma(Y)$ is generated by the π-system $\{\{Y \le y\}, y \in \mathbb{R}\}$. They are independent if and only if

$$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y)$$

which is equivalent to the desired equation.

(2) Consider $B = (-\infty, x] \times (-\infty, y]$. We have

$$\begin{aligned}
\mu_{X,Y}(B) &= \mu_{X,Y}((-\infty, x] \times (-\infty, y]) \\
&= F_{X,Y}(x,y) = F_X(x)F_Y(y) \\
&= \int_{-\infty}^{x} \int_{-\infty}^{y} f(u)g(v)dudv \\
&= \int_B f(u)g(v)\ dudv
\end{aligned}$$

as required. The equality holds for all B, which forms a π-system, so by Theorem 2.4, the equality holds on the whole of $\mathcal{B}(\mathbb{R}^2)$.

(3) We have

$$\begin{aligned}
F_{X+Y}(t) &= \mathbb{P}(X+Y \leq t) \\
&= \int_{-\infty}^{t}\int_{-\infty}^{\infty} P(X=s)\mathbb{P}(Y=u-s)dsdu \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{t} P(X=s)\mathbb{P}(Y=u-s)duds \\
&= \int_{-\infty}^{t}\int_{-\infty}^{\infty} P(X=s)\mathbb{P}(Y=u-s)dsdu.
\end{aligned}$$

Taking derivative on both sides yields

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} P(X=s)\mathbb{P}(Y=t-s)ds = \int_{\mathbb{R}} f(t)g(t-s)ds,$$

as desired. □

2.5 Infinitely Often and Eventually

Let $(\Omega, \Sigma, \mathbb{P})$ be the probability space we are working in.

Definition 2.10. *Let $\{E_n\}$ be a sequence of events in Σ. We have*

$$\{E_n \ i.o.^3\} = \bigcap_{N=1}^{\infty}\bigcup_{n\geq N} E_n$$

and

$$\{E_n \ ev.^4\} = \bigcup_{N=1}^{\infty}\bigcap_{n\geq N} E_n.$$

Notice that based on the definition, we have

$$\{E_n^C \ i.o.\}^C = \left[\bigcap_{N=1}^{\infty}\bigcup_{n\geq N} E_n^C\right]^C = \bigcup_{N=1}^{\infty}\bigcap_{n\geq N} E_n = \{E_n \ ev.\}.$$

So, any results about one of the two things can be easily transformed to be about the other.

The reason why we bring up these concepts is to better deal with strong convergence, i.e. $X_n \to X$ a.s. This is the same as saying the event $\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}$ has probability 1. We could

[3] infinitely often

[4] eventually

simply have

$$\begin{aligned}\{X_n \to X\} &= \{\forall \varepsilon, \exists N \in \mathbb{N}, \forall n \geq N \; |X_n - X| \leq \varepsilon\} \\ &= \{\forall k \in \mathbb{N}, \exists N \in \mathbb{N}, \forall n \geq N \; |X_n - X| \leq 1/k\} \\ &= \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|X_n - X| \leq 1/k\} \\ &= \bigcap_{k=1}^{\infty} \{|X_n - X| \leq 1/k \; ev.\}.\end{aligned}$$

Clearly, the events $\{|X_n - X| \leq 1/k \; ev.\}$ are decaying in k. So, to prove $\mathbb{P}(X_n \to X) = 1$ we just need to show that

$$\mathbb{P}(|X_n - X| \leq 1/k \; ev.) = 1$$

or equivalently

$$\mathbb{P}(|X_n - X| \geq 1/k \; i.o.) = 0.$$

Two lemmas are needed to get a lot of nice results. Those two lemmas are the Borel-Cantelli Lemma 1 and 2.

Lemma 2.11 (Borel-Cantelli Lemma). *We have*

1. *If* $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$, *then* $\mathbb{P}(E_n \; i.o.) = 0$.
2. *If* $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$ *and* $E_n s$ *are independent, then* $\mathbb{P}(E_n \; i.o.) = 1$.

Proof. (1) We want:

$$\mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n \geq N} E_n\right) = 0 \iff \lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n \geq N} E_n\right) = 0$$

since $\bigcup_{n \geq N} E_n$ is decreasing.

Notice that we have

$$\mathbb{P}\left(\bigcup_{n \geq N} E_n\right) \leq \sum_{n=N}^{\infty} \mathbb{P}(E_n)$$

which goes to 0 as $N \to \infty$ since $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$. So, by sandwich theorem, we have

$$\lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n \geq N} E_n\right) = 0$$

as desired.

(2) We want:

$$\mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n \geq N} E_n\right) = 1 \iff \mathbb{P}\left(\bigcup_{n \geq N} E_n\right) = 1 \; \forall N \iff \mathbb{P}\left(\bigcap_{n \geq N} E_n^C\right) = 0 \; \forall N$$

since $\bigcup_{n \geq N} E_n$ is decreasing.

Notice that we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{n\geq N} E_n^C\right) &= \lim_{M\to\infty}\mathbb{P}\left(\bigcap_{n=N}^{M} E_n^C\right) = \lim_{M\to\infty}\prod_{n=N}^{M}\mathbb{P}\left(E_n^C\right)\\ &= \lim_{M\to\infty}\prod_{n=N}^{M}[1-\mathbb{P}(E_n)] \leq \lim_{M\to\infty}\prod_{n=N}^{M}\exp[-\mathbb{P}(E_n)]\\ &= \lim_{M\to\infty}\exp\left[-\sum_{n=N}^{M}\mathbb{P}(E_n)\right] = \exp\left[-\lim_{M\to\infty}\sum_{n=N}^{M}\mathbb{P}(E_n)\right]\\ &= \exp\left[-\lim_{M\to\infty}\sum_{n=N}^{M}\mathbb{P}(E_n)\right] = \exp\left[-\infty\right] = 0\end{aligned}$$

where the inequality is due to the fact that $1-x\leq e^{-x}$ for all x and the second equality is by the independence of E_ns. □

–

One thing we can do with BC is to study the extreme values of a sequence of i.i.d. random variables.

Let $\{X_n\}$ be i.i.d. sequence of Exp(1), such that for any n we have

$$\mathbb{P}(X_n > x) = e^{-x} \quad x>0.$$

Consequently, for any $\alpha>0$, we have

$$\mathbb{P}(X_n > \alpha\log n) = n^{-\alpha}.$$

We let $L_n := X_n/\log n$. We claim that with probability one we have $\limsup L_n = 1$. To show this result, we need to establish that (1) for any $k>1$, $L_n > k$ only finitely many times, and (2) for any $k<1$, $L_n>k$ infinitely often.

To show the first result, for $k>1$, we have

$$\sum_n \mathbb{P}(L_n > k) = \sum_n \mathbb{P}(X_n > k\log n) = \sum_n n^{-k} < \infty$$

so BC1 says that $L_n > k$ finitely often with probability one.

To show the second result, for $k<1$, we have

$$\sum_n \mathbb{P}(L_n > k) = \sum_n \mathbb{P}(X_n > k\log n) = \sum_n n^{-k} = \infty$$

so BC2 says that $L_n > k$ infinitely often.

Thus, we have established that $\limsup X_n/\log n = 1$.

The above is a special case, and now let us look at a more general result.

Let $\{X_n\}$ be a sequence of i.i.d. random variables. With probability one, we have

$$\limsup_{n\to\infty}\frac{|X_n|}{n} = \begin{cases}0 & \text{if } \mathbb{E}[|X_1|]<\infty\\ \infty & \text{if } \mathbb{E}[|X_1|]=\infty\end{cases}$$

The first case follows straight from the SLLN[5]. We have

$$\frac{X_n}{n}=\frac{X_1+\cdots+X_n}{n}-\frac{X_1+\cdot+X_{n-1}}{n-1}\cdot\frac{n-1}{n}\to\mathbb{E}[X_1]-\mathbb{E}[X_1]=0\quad a.s.$$

In the second case, we cannot use SLLN since the expectation is infinity. We want

$$\begin{aligned}\mathbb{P}\left(\limsup_{n\to\infty}\frac{|X_n|}{n}=\infty\right)&=1\\ \mathbb{P}\left(\bigcap_{m=1}^{\infty}\{|X_n|/n>m\ i.o.\}\right)&=1\\ \mathbb{P}\left(\{|X_n|/n>m\ i.o.\}\right)&=1\quad\text{for all } m.\end{aligned}$$

For each m, the events $\{|X_n|/n>m\}$ for $n\in\mathbb{N}$ are independent and we have

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbb{P}(|X_n|/n>m)=\sum_{n=1}^{\infty}\mathbb{P}(|X_1|/n>m)&=\sum_{n=1}^{\infty}\mathbb{E}(1_{\{|X_1|/n>m\}})\\ &=\mathbb{E}\left[\sum_{n=1}^{\infty}1_{\{|X_1|/n>m\}}\right]\quad\text{MON}\\ &=\mathbb{E}\left[\sum_{n=1}^{\infty}1_{\{|X_1|/m>n\}}\right]\geq\mathbb{E}\left[\frac{|X_1|}{m}-1\right]=\infty.\end{aligned}$$

The rest follows from BC.

We will state without proof the following theorem that is an advancement of the above results.

Theorem 2.12 (Law of Iterated Logarithm)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean 0 and variance 1. Let us then denote $S_n:=X_1+\cdots+X_n$. We have,*

$$\limsup_{n\to\infty}\frac{|S_n|}{\sqrt{2n\log\log n}}=1\quad a.s.$$

–

Next, we will try to prove the strong law of large number (SLLN) under strong conditions. The proof requires the Borel-Cantelli Lemma as well as the Bernstein Inequality. Let us prove this new inequality first.

Theorem 2.13 (Bernstein Inequality)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables with $\mathbb{P}(X_i=1)=\mathbb{P}(X_i=-1)=1/2$ for all i. Then, for any $a_1,a_2,\ldots\in\mathbb{R}$ that are not all zero, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}a_iX_i\right|\geq t\right)\leq 2\exp\left[-\frac{t^2}{2\sum_{i=1}^{n}a_i^2}\right].$$

Proof. We let $c=\sum_{i=1}^{n}a_i^2$ and consider some $\lambda>0$. Then, we have

$$\mathbb{E}\exp\left[\lambda\sum_{i=1}^{n}a_iX_i\right]=\mathbb{E}\prod_{i=1}^{n}e^{\lambda a_iX_i}=\prod_{i=1}^{n}\mathbb{E}e^{\lambda a_iX_i}.$$

[5] SLLN states that for an i.i.d. sequence with finite expectation, their means converge to the expectation almost surely. Various forms of this result will be proved throughout this notes.

Additionally, we have

$$Ee^{\lambda a_i X_i} = \frac{1}{2}[e^{\lambda a_i} + e^{-\lambda a_i}]$$

and

$$[e^{\lambda a_i/2}]^2 + [e^{-\lambda a_i/2}]^2 \leq 2.$$

Note that we have

$$(e^x + e^{-x})/2 = \frac{1}{2}\sum_{n=1}^{\infty} \frac{x^n + (-x)^n}{n!} = \sum_{k=1}^{\infty} \frac{x^{2k}}{(2k)!} \leq \sum_{k=1}^{\infty} \frac{x^{2k}}{2^k \cdot k!} = \sum_{k=1}^{\infty} \frac{(x^2/2)^k}{k!} = e^{x^2/2}.$$

So, we have

$$\prod_{i=1}^{n} \mathbb{E}e^{\lambda a_i X_i} \geq \prod_{i=1}^{n} e^{\lambda^2 a_i^2/2} = e^{\lambda^2 \sum a_i^2/2} = e^{\lambda^2 c/2}.$$

Using the Markov inequality, we get

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i \geq t\right] = \mathbb{P}\left[\exp\{\lambda \sum a_i X_i\} \geq e^{\lambda t}\right] \leq \mathbb{E}\exp\{\lambda \sum a_i X_i\} \cdot e^{-\lambda t}) \leq e^{\lambda^2 c/2 - \lambda t}.$$

This holds for all λ, so pick the value that minimises the upper bound, which is $\lambda = t/c$. Thus, we get

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i \geq t\right] \leq e^{-t^2/(2c)}$$

and similarly

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i \leq -t\right] \leq e^{-t^2/(2c)}$$

which gives us the desired result after combining them. □

This helps us to obtain the SLLN under strong conditions.

Theorem 2.14. *Let $\{X_n\}$ be a sequence of i.i.d. random variables with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$ for all i. Then, we have*

$$\frac{X_1 + \cdots + X_n}{n} \to 0 \qquad a.s.$$

Proof. As discussed earlier, the desired result is equivalent to showing that

$$\mathbb{P}\left[\left|\frac{X_1 + \cdots + X_n}{n}\right| \geq \frac{1}{k} \quad \text{i.o.}\right] = 0$$

and

$$\mathbb{P}\left[|X_1 + \cdots + X_n| \geq \frac{n}{k} \quad \text{i.o.}\right] = 0.$$

Using the Borel-Cantelli Lemma, we just need to show $\sum_{n=1}^{\infty} \mathbb{P}(|\sum X_i| \geq n/k) < \infty$. Using the Bernstein inequality, we have

$$\mathbb{P}(\left|\sum X_i\right| \geq n/k) \leq 2\exp\left[-\frac{n^2/k^2}{2n}\right] = 2\exp\left[-\frac{n}{2k^2}\right].$$

So, we get

$$\sum_{n=1}^{\infty} \mathbb{P}(\Big|\sum X_i\Big| \geq n/k) = \sum_{n=1}^{\infty} 2\exp\Big[-\frac{n}{2k^2}\Big] = 2\sum_{n=1}^{\infty} \big[\exp[-1/(2k^2)]\big]^n < \infty$$

as the exponential is between 0 and 1. This yields the desired result. □

2.6 Tail Events and Kolmogorov 0-1 Law

Definition 2.15. *Let $\{X_n\}$ be a sequence of random variables. $\tau_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$ is the n-**th tail σ-algebra** of $\{X_n\}$. $\tau := \cap_{n=1}^{\infty} \tau_n$ is the **tail σ-algebra** of $\{X_n\}$. An event is a **tail event** if $E \in \tau$.*

The intuition behind this definition is as follows. $\sigma(X) = \{\{X \in B\} \mid B \in \mathcal{B}\}$, which contains all events that depend on X. So, τ_n contains events which are defined by $X_{n+1}, X_{n+2}, \ldots$ only, or events that do not depend on $X_1, \ldots, X_n$. Extrapolating this intuition, we have that τ contains events which are determined by the tail of $\{X_n\}$, or events that do not depend on any finite number of X_n.

Example. $E = \{X_n \to a\}$.

If this is a tail event, then it means it belongs to the intersection of τ_m so it must be contained in any τ_m. To see this, for any m, we have

$$\{X_n \to a\} = \{X_{n+m} \to a\} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n \geq N} \{|X_{m+n-a}| < \frac{1}{k}\} \in \tau_m.$$

So this is indeed a tail event.

Example. $\{\lim_{n\to\infty} X_n \text{ exists}\} \in \tau$.

Example. $\{\sum_{n=1}^{\infty} X_n < \infty\} \in \tau$.

Example. $\{\sum_{n=1}^{\infty} X_n > 0\} \notin \tau$.

Example. $\{\lim_{n\to\infty} [X_1 + \cdots + X_n]/n \text{ exists}\} \in \tau$.

To see this, notice that $[X_1 + \cdots + X_n]/n = [X_1 + \cdots + X_m]/n + [X_{m+1} + \cdots + X_n]/n$ and the first term on the right goes to zero as $n \to \infty$. So any finite term will not matter.

Example. $\{\sup_{n\in\mathbb{N}} X_n > 0\} \notin \tau$.

Assume that it is a tail event and find a counter-example.

Consider the sequence of random variables such that X_1 takes 0 and 2 each with probability 1/2, and $X_n = 0$ for all $n \geq 2$. Then, $E := \{\sup_{n\in\mathbb{N}} X_n > 0\} = \{X_1 = 2\}$ and $\mathbb{P}(E) = 1/2$. Furthermore, $\tau_n = \{\emptyset, \Omega\}$ for all $n \geq 2$ so $\tau = \{\emptyset, \Omega\}$. Clearly, $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$, and neither is $\mathbb{P}(E) = 1/2$. So, this is not a tail event.

Example.

$$\left\{ \limsup \frac{1}{n} \sum_{i=1}^{n} X_i > 0 \right\}$$

is a tail event.

Notice that we have, for any m,

$$\limsup \frac{1}{n} \sum_{i=1}^{n} X_i = \limsup \frac{1}{n} \sum_{i=m+1}^{n} X_i$$

as the limit of the first finite many m terms go to zero by the $1/n$ factor.

Example.

$$\left\{ \lim \frac{1}{n} \sum_{i=1}^{n} X_i > 0 \ i.o. \right\}$$

is not a tail event.

For example, if we consider i.i.d. $X_1 \sim Ber(1/2)$ and $X_n = 0$ for $n \geq 2$, then the above event depends exactly on the behaviour of the first term.

Theorem 2.16 (Kolmogorov 0-1 Law)**.** *If $\{X_n\}$ are independent, then every tail event has a probability of 0 or 1.*

Proof. We will show that τ is independent of τ (which is only possible when τ is trivial). This then imply, for all $E \in \tau$, we have $\mathbb{P}(E \cap E) = \mathbb{P}(E)\mathbb{P}(E) \implies \mathbb{P}(E) = \mathbb{P}(E)^2 \implies \mathbb{P}(E) = 0, 1$, as desired.

Let $\sigma_n := \sigma(X_1, X_2, \ldots, X_n)$. Recall $\tau_n := \sigma(X_{n+1}, X_{n+2}, \ldots)$.

The proof will be in steps.

(a) $\sigma_n \perp \tau_n$.

We have two π-systems,

$$\{\{X_1 \in B_1, \ldots, X_n \in B_n\}, B_1, \ldots, B_n \in \mathcal{B}\}$$

that generates σ_n and

$$\{\{X_{n+1} \in B_{n+1}, \ldots, X_{n+m} \in B_{n+m}\}, B_{n+1}, \ldots, B_{n+m} \in \mathcal{B}\}$$

that generates τ_n. Clearly, they are independent, so by Theorem 2.3, they are indeed independent.

(b) Since $\tau \subset \tau_n$, we have $\sigma_n \perp \tau$.

(c) Let $\sigma_\infty := \sigma(X_1, X_2, \ldots)$. Then, $\sigma_\infty \perp \tau$.

σ_∞ is generated by π-system $\sigma_1 \cup \sigma_2 \cup \cdots$. First, this is indeed a π-system. For $A, B \in \sigma_1 \cup \sigma_2 \cup \cdots$, we have $A \in \sigma_m$ and $B \in \sigma_n$. Since σ_ks are increasing, we have $A, B \in \sigma_{\max\{m,n\}}$ and also $A \cap B \in \sigma_{\max\{m,n\}}$. So, we have $A, B \in \sigma_1 \cup \sigma_2 \cup \cdots$. Second, $\sigma_1 \cup \sigma_2 \cup \cdots \perp \tau$. This is true as $\sigma_n \perp \tau$ for all n. Thus, $\sigma_\infty \perp \tau$.

(d) $\tau \subset \sigma_\infty$.

Thus, $\tau \perp \tau$.

□

At this point, we should go back to one of the result we showed earlier again. We said that for i.i.d. exponential random variables X_n with mean 1, we have

$$\limsup \frac{X_n}{\log n} = 1.$$

We assumed that this limsup is a constant a.s. but it could be verified now using Kolmogorov's 0-1 law.

Let F be the distribution function of $\limsup X_n$. Then $F(x) = \mathbb{P}(\limsup X_n \leq x)$. The event $\{\limsup X_n \leq x\}$ is a tail event so its probability is either 0 or 1. Hence F can only take values 0 and 1 - which is only possible if the underlying random variable is a constant.

Chapter 3

Weak Convergence

3.1 Basics of Weak Convergence

Let $\{X_n\}$, X be random variables, and they do not have to be on the same probability space. Respectively, let μ_n, μ be their laws and let F_n, F be their distribution functions.

Definition 3.1. *X_n converges to X **weakly / in distribution / in law**, or $\mu_n \to \mu$ weakly, if $F_n(t) \to F(t)$ for every $t \in \mathbb{R}$ where F is continuous. We will denote this by $X_n \implies X$, or $X_n \xrightarrow{d} X$, $X_n \xrightarrow{w} X$.*

Theorem 3.2 (WLLN for square-integrable random variables)**.** *Let $\{X_n\}$ be i.i.d. random variables with mean μ and finite variance. Then,*

$$\frac{X_1 + \cdots + X_n}{n} \implies \mu.$$

Proof. WLOG we let μ to be zero. Also, we denote $S_n := X_1 + \cdots + X_n$.

Note that for some variable Y to have $\mathbb{P}(Y = 0) = 1$, it must have distribution function

$$F_Y(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0. \end{cases}$$

So, in this case, for any $\varepsilon > 0$, we need $F_{S_n/n}(\varepsilon) \to 1$ and $F_{S_n/n}(-\varepsilon) \to 0$ as $n \to 0$. Essentially, we just need $\mathbb{P}(|S_n/n| > \varepsilon) \to 0$ as $n \to 0$.

Using Markov inequality, we have

$$\begin{aligned} \mathbb{P}(|S_n/n| > \varepsilon) &= \mathbb{P}((S_n/n)^2 > \varepsilon^2) \\ &= \mathbb{P}(S_n^2 > n^2\varepsilon^2) \\ &\leq n^{-2}\varepsilon^{-2}\mathbb{E}[S_n^2] \\ &= n^{-2}\varepsilon^{-2}\mathbb{E}[\sum_n X_n^2] \quad \text{as } X_n \text{ are independent} \\ &= \frac{n\mathrm{Var}(X_1)}{n^2\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\varepsilon^2} \to 0. \end{aligned}$$

□

Theorem 3.3 (Relation Between Weak and Almost Surely Convergence)**.** *We have*

1. *If $X_n \to X$ a.s., then $X_n \to X$ weakly.*
2. *If $\mu_n \to \mu$ weakly, then there are $\{X_n\}$, X defined on the same probability space such that their laws are $\{\mu_n\}$, μ, and $X_n \to X$ a.s.*

Theorem 3.4 (Equivalent Definition of Weak Convergence 1)**.** *$\mu_n \to \mu$ weakly $\Longleftrightarrow \int_{\mathbb{R}} h \; d\mu_n \to \int_{\mathbb{R}} h \; d\mu$ for all continuous and bounded function $h : \mathbb{R} \to \mathbb{R}$.*

Proofs of the two theorems above will be combined, as they are intertwined.

Proof. We will first prove Theorem 3.3 (2).

We will use the Skorokhod's construction (see Theorem 1.5) to construct the random variables X_n and X. Let $([0,1), \mathcal{B}, Leb)$ be the probability space, and we define random variables

$$X_n(\omega) := \inf\{t : F_n(t) > \omega\}$$

and

$$X(\omega) := \inf\{t : F(t) > \omega\}$$

where F_n and F are the distribution functions from the laws μ_n and μ.

Let $B := \{\omega \in [0,1) : F(x) = F(y) = \omega$ for some $x \neq y\}$. This defines the set of constant pieces of F. We will show in a bit that B has Lebesgue measure zero (and in fact it is a countable set), so we will disregard them in checking the strong convergence of X_n to X.

For each $\omega \in B$, we will associate it with an interval $[x, y]$ such that $F(x) = F(y) = \omega$. This is a bijection. Furthermore, the intervals for two different ω are trivially disjoint. Notice that each interval contains a rational number, and the set of rational numbers is countable, so we have at most countable such intervals, and therefore B is at most countable. This implies that $Leb(B) = 0$.

Consider the set of discontinuities of F. Using a similar argument, we can establish that there are at most countable discontinuities. For each discontinuity $t \in \mathbb{R}$, we can associate it with interval $[v, u]$ such that $\lim_{x \to t^-} F(x) \leq v < u \leq lim_{x \to t^+} F(x)$. Those vertical intervals are disjoint, and each contains a rational number, so there could only be at most countable many of them - thus there are at most countable discontinuities.

Let $\omega \in [0,1) \backslash B$, and consider some $\varepsilon > 0$ with $0 < \delta \leq \varepsilon$ such that for $x = X(\omega)$, we have $x - \delta$ and $x + \delta$ both being continuity points of F. This is possible since there are at most countable discontinuity points of F, as established just now. So, we have

$$F(x - \delta) < \omega < F(x + \delta)$$

with $F_n(x - \delta) \to F(x - \delta)$ and $F_n(x + \delta) \to F(x + \delta)$ by the given weak convergence $\mu_n \implies \mu$. Some more, as we $\omega \notin B$, we can find sufficiently large N such that for all $n \geq N$, we have $F_n(x - \delta) < \omega < F_n(x + \delta)$.

As a consequence of the above inequality, we have $x - \delta < X_n(\omega) < x + \delta$, which yields

$$|X(\omega) - X_n(\omega)| < \delta$$

for all $n \geq N$, as required for $X_n \to X$ a.s..

We will then prove the forward direction of Theorem 3.4.

As proven just now, given $\mu_n \implies \mu$, we can construct random variables X_n and X such that $X_n \to X$ a.s. Next, notice that the desired convergence is

$$\mathbb{E}[h(X_n)] = \int h d\mu_n \to \int h d\mu = \mathbb{E}[h(X)].$$

Since h is continuous, we have $X_n \to X$ *a.s.* $\implies h(X_n) \to X$ *a.s.*. In addition, h is bounded so $h(X_n)$ is bounded. Therefore, we apply the dominated convergence theorem and get

$$\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)] \ a.s.$$

as desired.

Next, we will prove the backward direction of the same theorem.

Showing $\mu_n \implies \mu$ is the same as showing $F_n(x) \to F(x)$ for all continuous point x. Notice that

$$F_n(x) = \int 1_{(-\infty,x]} d\mu_n$$

and

$$F(x) = \int 1_{(-\infty,x]} d\mu.$$

We would be tempted to draw the convergence using the condition of the theorem $\int_{\mathbb{R}} h \ d\mu_n \to \int_{\mathbb{R}} h \ d\mu$ for $h = 1_{(-\infty,x]}$. However, this is not valid as h is bounded but not continuous. We could remedy this argument by trying to make the indicator function continuous, and that is what we will do.

Consider a continuity point x of F. The desired convergence is $\lim_n F_n(x) = F(x)$, and this is implied by

$$F(x) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x).$$

The middle inequality is trivial, and we just need to show the first and last inequalities.

First we show $F(x) \leq \liminf F_n(x)$. For some $\delta > 0$, consider $h(t)$ that takes 1 for $t \leq x - \delta$, takes 0 for $t \geq x$, and something between 0 and 1 for $x - \delta \leq t \leq x$ that makes h continuous. Clearly, we have

$$\liminf F_n(x) = \liminf \int 1_{(-\infty,x]} d\mu_n \geq \liminf \int h d\mu_n = \int h d\mu \geq \int 1_{(-\infty,x-\delta]} d\mu = F(x-\delta).$$

As this δ is arbitrary, we can take it to zero and get the desired

$$\liminf F_n(x) \geq F(x).$$

The other inequality is similar. We just need to consider $h(t)$ that takes 1 for $t \leq x$, takes 0 for $t \geq x + \delta$, and something between 0 and 1 for $x \leq t \leq x + \delta$ that makes h continuous for any $\delta > 0$. The rest is almost identical, and we would have

$$\limsup F_n(x) \leq F(x).$$

Done.

Finally, we will prove Theorem 3.3 (1).

Using the backward direction of Theorem 3.4, we could get $\mu_n \to \mu$ weakly if we have $\int_{\mathbb{R}} h\ d\mu_n \to \int_{\mathbb{R}} h\ d\mu$ for all continuous and bounded function $h : \mathbb{R} \to \mathbb{R}$, and that implies the desired weak convergence $X_n \implies X$ using Theorem 3.3 (2).

We want

$$\mathbb{E}[h(X_n)] = \int_{\mathbb{R}} h\ d\mu_n \to \int_{\mathbb{R}} h\ d\mu = \mathbb{E}[h(X)].$$

Since $X_n \to X$ a.s., we have $h(X_n) \to h(X)$ a.s. by the continuity of h, and $h(X_n)$ is bounded by the boundedness of h. So, the desired convergence is established using the dominated convergence theorem. □

Definition 3.5. *$h : \mathbb{R} \to \mathbb{R}$ is a C^2-**test function** if it is compactly supported, twice differentiable, and h'' is continuous.*

Theorem 3.6 (Equivalent Definition of Weak Convergence 2)**.** *$\mu_n \to \mu$ weakly $\iff \int_{\mathbb{R}} h\ d\mu_n \to \int_{\mathbb{R}} h\ d\mu$ for all C^2-test function $h : \mathbb{R} \to \mathbb{R}$.*

Definition 3.7. *$\{\mu_n\}$ is **tight** if for all $\varepsilon > 0$, there exists M such that $\mu_n([-M, M]) \geq 1 - \varepsilon$ for all n.*

Example. Any finite collection is tight.

Example. $\mu_n = \delta_n$. Not tight. For some ε and corresponding M, we can find $mu_{M+1}[-M, M] = 0$ that breaks the condition.

Example. $\mu_n \sim N(a_n, 1)$. Tight if and only if $\{a_n\}$ is bounded.

Example. $\mu_n \sim N(0, \sigma_n^2)$. Tight if and only if $\{\sigma_n^2\}$ is bounded.

Lemma 3.8. *$\int_{\mathbb{R}} h\ d\mu_n \to \int_{\mathbb{R}} h\ d\mu$ for all C^2-test function $h : \mathbb{R} \to \mathbb{R}$ implies that $\{\mu_n\}$ is tight.*

Proof. Let $\varepsilon > 0$. Choose M_1 such that $\mu[-M_1, M_1] \geq 1 - \varepsilon/2$. Consider a C^2 test function h that takes 1 on $[-M_1, M_1]$, takes 0 outside $(-1 - M_1, 1 + M_1)$, and something between 0 and 1 otherwise so that h is indeed a C^2 test function.

Next, we have

$$\begin{aligned}\mu_n[-M_1, M_1] &= \int 1_{[-M_1, M_1]} d\mu_n \\ &\geq hd\mu_n \to \int hd\mu \text{ for large enough } n \\ &\geq \int hd\mu - \frac{\varepsilon}{2} \\ &\geq \mu[-M_1, M_1] - \frac{\varepsilon}{2} \geq 1 - \varepsilon.\end{aligned}$$

The tightness requires M for all n. We have so far shown that for M_1 and large enough n. There are only finitely many n that M_1 might not be large enough. So, we pick finitely many M_k and pick the maximum out of them, and that will settle the proof. □

Now, let us prove Theorem 3.6.

Proof. The forward direction is trivially true as a consequence of Theorem 3.4 since any h that is a C^2 test function is obviously continuous and bounded. So we just need to show the backward direction.

We know from the previous lemma that $\{\mu_n\}$ is tight. Let x be a continuity point of F, and let $\delta > 0$, $\varepsilon > 0$. For ε, we can pick M such that $\mu_n[-M, M] \geq 1-\varepsilon$ for all n, and $\mu[-M, M] \geq 1-\varepsilon$.

The desired convergence can be implied by

$$F(x) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x).$$

Just like our proof for the backward direction of Theorem 3.4, we will consider some suitable h to make the whole inequalities hold.

First, we show $F(x) \leq \liminf F_n(x)$. Consider h that takes 1 on $[-M, x-\delta]$, takes 0 from x onwards and before $-M-1$, and takes something between 0 and 1 otherwise to make sure h is indeed a C^2 test function.

We have $F_n(x) = M_n(-\infty, x] = \int 1_{(-\infty,x]} d\mu_n \geq \int h d\mu_n$. Then,

$$\begin{aligned}\liminf F_n(x) &\geq \liminf \int h d\mu_n = \int_{\mathbb{R}} h d\mu \\ &\geq 1_{[-M,x-\delta]} d\mu = F(x-\delta) - \mu(-\infty, -M] \\ &\geq F(x-\delta) - \varepsilon.\end{aligned}$$

As δ, ε are arbitrary, we take them to zero and have $\liminf F_n(x) \geq F(x)$. The other inequality is similar. Just consider a C^2 test function h that takes 0 on $[-M, x]$, takes 1 outside $(-M-1, x+\delta)$, and something nice in between. The rest follows similarly. Done. □

3.2 Characteristic Functions

Definition 3.9. *We have*

1. $\varphi(t) = \mathbb{E}[e^{itX}]$, $\varphi : \mathbb{R} \to \mathbb{C}$ *is the* ***characteristic function*** *of random variable* X.
2. $\hat{\mu}(t) = \int e^{itx} d\mu(x)$, $\hat{\mu} : \mathbb{R} \to \mathbb{C}$ *is the* ***Fourier transform*** *of law* μ.
3. $\hat{f}(t) = \int e^{itx} f(x) dx$, $\hat{f} : \mathbb{R} \to \mathbb{C}$ *is the* ***Fourier transform*** *of integrable function* $f : \mathbb{R} \to \mathbb{R}$.

Theorem 3.10 (Properties of Characteristic Functions)**.** *For a characteristic function* ϕ*, we have*

1. $\varphi(0) = 1$.
2. $\varphi_{\lambda X}(t) = \varphi_X(\lambda t)$.
3. *If* X, Y *are independent, then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\phi_Y(t).$$

4. φ *is continuous.*

Proof. The first result follows from the definition of φ. The second result is obtained by seeing

$$\varphi_{\lambda X}(t) = \mathbb{E}[e^{it\lambda X}] = \varphi_X(\lambda t).$$

The third result follows from

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{it\lambda(X+Y)}] = \mathbb{E}[e^{it\lambda X}]\mathbb{E}[e^{it\lambda Y}] = \varphi_X(t)\varphi_Y(t),$$

where the second last equality follows from the independence of X and Y. The last result is true by considering any $t_n \to t$, and we have

$$\varphi(t_n) = \mathbb{E}e^{it_n x} \to \mathbb{E}e^{itx} = \varphi(t)$$

by the continuity of exponential and expectation. □

Example. For a random variable X that takes a constant value c with probability 1, we have

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = e^{itc}.$$

Example. For a random variable X that takes ±1 each with probability 1/2, we have

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t.$$

Example. Consider $X \sim Unif[-1, 1]$. We have

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_{-1}^{1} \frac{1}{2}e^{itx}dx = \frac{1}{2it}(e^{it} - e^{-it}) = \frac{\sin t}{t}$$

for $t \neq 0$. For $t = 0$, we have $\varphi(0) = 1$ and the continuity of ϕ is maintained.

Example. For Cauchy random variable X with pdf $f(x) = 1/[\pi(1 + x^2)]$, it has infinite expectation. Additionally, it can be derived using some complex analysis that

$$\varphi(t) = e^{-t|x|}.$$

This function is not differentiable. We have this observation that if for some random variable X we have $\mathbb{E}[|X|^m] < \infty$, then ϕ will be m-times differentiable.

Example. For $X \sim N(0, 1)$, we have

$$\varphi(t) = e^{-t^2/2}$$

using complex analysis.

Theorem 3.11 (Decay and Integrability of Fourier Transform)**.** *We have*

1. *For $h : \mathbb{R} \to \mathbb{R}$, if we have*
   - *h is m-times differentiable and $h^{(m)}$ is continuous,*
   - *$h^{(i)}$ is integrable for $0 \leq i \leq m$,*
   - *$h^{(i)}(x) \to 0$ as $x \to \infty$ for all $0 \leq i \leq m - 1$,*

   then there exits some c such that $|\hat{h}(t) \leq C/t^m$ for all non-zero t.
2. *Let h be a C^2 test function, then there some c such that $|\hat{h}(t) \leq C/t^2$ for all non-zero t.*

Proof. The second part of this theorem is a direct consequence of the first part after noticing that $m = 2$ for a C^2 test function.

Using integration by parts, we have

$$\begin{aligned}\hat{h}(t) &= \int_{\mathbb{R}} e^{itx} h(x) dx \\ &= \left[\frac{h(x)e^{itx}}{it}\right]_{\mathbb{R}} - \frac{1}{it}\int_{\mathbb{R}} h'(x)e^{itx}dx \\ &= -\frac{1}{it}\int_{\mathbb{R}} h'(x)e^{itx}dx\end{aligned}$$

as h decays to zero for large $|x|$. We can keep doing this integration by parts $m-1$ more times, and that will give us

$$\hat{h}(t) = \frac{(-1)^m}{(it)^m}\int_{\mathbb{R}} h^{(m)}(x)e^{itx}dx,$$

which implies

$$|\hat{h}(t)| = \left|\frac{(-1)^m}{(it)^m}\int_{\mathbb{R}} h^{(m)}(x)e^{itx}dx\right| \leq \frac{1}{t^m}\int |h^{(m)}(x)e^{itx}|dx = \frac{c}{t^m}$$

for some constant c and $t \neq 0$. □

Theorem 3.12 (Parseval-Plancherel Theorem)**.** *Let μ be a probability measure on $(\mathbb{R}, \Omega)$ with Fourier transform φ. Then, for any C^2 test function, we have*

$$\int_{\mathbb{R}} h d\mu = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{\hat{h}(t)}\varphi(t)dt.$$

Proof. The expression on the RHS is integrable since $|\varphi| < 1$ and $|\overline{\hat{h}}| \leq c/t^2$ as we have just established in Theorem 3.11. So, we can use Fubini and have

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{\hat{h}(t)}\varphi(t)dt &= \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\overline{\hat{h}(t)}\varphi(t)dt \\ &= \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\int_{-\infty}^{\infty} e^{-itx}h(x)dx\int_{\mathbb{R}}\int e^{ity}d\mu(y)dt \\ &= \frac{1}{2\pi}\lim_{T\to\infty}\int_{\mathbb{R}}\int_{-\infty}^{\infty} h(x)\int_{-T}^{T} e^{it(y-x)}dtdxd\mu(y) \\ &= \lim_{T\to\infty}\int_{\mathbb{R}}\frac{1}{\pi}\int_{-\infty}^{\infty} h(x)\frac{\sin(T(x-y))}{x-y}dxd\mu(y).\end{aligned}$$

We will denote $h^T(y) := \frac{1}{\pi}\int_{-\infty}^{\infty} h(x)\frac{\sin(T(x-y))}{x-y}dx$.

In order to show the desired result

$$\int_{\mathbb{R}} h d\mu = \lim_{T\to\infty}\int_{\mathbb{R}} h^T(y)d\mu,$$

we need two things: (1) $h^T(y) \to h(y)$ as $T \to \infty$, and (2) h^T are uniformly bounded by an integrable function. If we have these things, then a simple application of the dominated convergence theorem will yield the desired equation.

We will write $h^T = h^T_- + h^T_+$, where we have

$$h^T_-(y) := \frac{1}{\pi}\int_{-\infty}^{y} h(x)\frac{\sin(T(x-y))}{x-y}dx$$

$$h^T_+(y) := \frac{1}{\pi}\int_{y}^{\infty} h(x)\frac{\sin(T(x-y))}{x-y}dx.$$

We just need to show $h^T_+(y) \to h(y)/2$. The other convergence can be established similarly. By integration by parts, we have

$$\begin{aligned} h^T_+(y) &= \frac{1}{\pi}\int_{y}^{\infty} h(x)\frac{\sin(T(x-y))}{x-y} \\ &= \frac{1}{\pi}\left(h(x)\int_y^x \frac{\sin(T(u-y))}{u-y}du|_{x=y}^{\infty} - \int_y^\infty h'(x) int_y^x \frac{\sin(T(u-y))}{u-y}dudx\right) \\ &= -\frac{1}{\pi}\int_y^\infty h'(x)\int_0^{T(x-y)}\frac{\sin v}{v}dvdx \end{aligned}$$

where we used the substitution $v = (u-y)/T$.

We claim that $\int_0^\infty (\sin t)/tdt = \pi/2$ and the function $x \mapsto \int_0^x (\sin t)/tdt$ is bounded by a positive constant.

Using this claim, for each x and as $T \to \infty$, we have the inner integral converging to $\pi/2$. So, using dominated convergence theorem since h' is compactly supported, we have

$$h^T_+(y) = -\frac{1}{\pi}\int_y^\infty h'(x)\int_0^{T(x-y)}\frac{\sin v}{v}dvdx \to -\frac{1}{\pi}\int_y^\infty h'(x)\pi/2dx = h(y)/2.$$

Next, we need to show h^T_+ is uniformly bounded, and this can be established by our previous remark. Done. □

Theorem 3.13 (Weak Convergence = Convergence of Characteristic Functions)**.** *Let $\{X_n\}$ and X be random variables with characteristic functions (φ_n) and φ. Then $X_n \implies X$ if and only if $\varphi_n \to \varphi$ for each $t \in \mathbb{R}$.*

Proof. We will let (μ_n) and μ to denote the laws of $\{X_n\}$ and X respectively.

We first show the forward direction. For each t, we have $h(x) := e^{itx}$ and this function is continuous and bounded. By Theorem 3.4, we know that if $\mu_n \implies \mu$ then $\int hd\mu_n \to \int hd\mu$ for every continuous bounded h. So, as we have $X_n \implies X$ so $\mu_n \implies \mu$, we have

$$\varphi_n(t) = \mathbb{E}[e^{itX_n}] = \int h(x)d\mu_n \to \int h(x)d\mu = \mathbb{E}[e^{itX}] = \varphi(t)$$

as desired.

For the backward direction, we will use Theorem 3.6, which says that $\mu_n \implies \mu$ if and only if $\int hd\mu_n \to \int hd\mu$ for every C^2 test function h. So, we just need to show that $\int hd\mu_n \to \int hd\mu$ for every C^2 test function h.

From Parseval-Plancherel (Theorem 3.12), we know that for any C^2 test function h, we have

$$\int_{\mathbb{R}} hd\mu = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{\hat{h}(t)}\varphi(t)dt.$$

So, we have

$$\int_{\mathbb{R}} hd\mu_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{\hat{h}(t)}\varphi_n(t)dt \to \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{\hat{h}(t)}\varphi(t)dt \int_{\mathbb{R}} hd\mu,$$

where the convergence holds due to dominated convergence theorem and Theorem 3.11 on the bound of text functions. Done. □

Example. Let $\{X_n\}$ be i.i.d. sequence taking ±1 each with probability 1/2. We would like to investigate the convergence of $\sum_{n=1}^{\infty} 2^{-n}X_n$.

We let $S_n := \sum_{k=1}^{n} 2^{-k}X_k$. We have

$$\phi_{S_n}(t) = \prod_{k=1}^{n} \phi_{X_k}(t2^{-k}) = \prod_{k=1}^{n} \cos(t2^{-k}).$$

Then,

$$\begin{aligned}
\phi_{S_n}(t) &= \frac{1}{\sin(t2^{-n})}\prod_{k=1}^{n}\cos(t2^{-k})\sin(t2^{-n}) \\
&= \frac{2^{-1}}{\sin(t2^{-n})}\prod_{k=1}^{n-1}\cos(t2^{-k})\sin(t2^{-(n-1)}) \\
&= \frac{2^{-n}}{\sin(t2^{-n})}\sin(t) \\
&= \frac{\sin(t)}{t}\frac{t2^{-n}}{\sin(t2^{-n})} \to \frac{\sin(t)}{t}
\end{aligned}$$

which is the characteristic function for Unif[-1,1]. Thus, this series converges to Unif[-1,1] in distribution.

So far we have obtained a series of equivalence relationships between convergences. They are summarised in the following diagram.

$$\begin{tikzcd}
 & \boxed{\begin{array}{l}\int hd\mu_n \to \int hd\mu \\ \text{for continuous and bounded } h\end{array}} \arrow[drr, Rightarrow] & & \\
\boxed{X_n \Rightarrow X} \arrow[ur, Leftrightarrow] \arrow[dr, Leftrightarrow] & & & \boxed{\varphi_n(t) \to \varphi(t) \text{ for all } t} \arrow[dl, Rightarrow] \\
 & \boxed{\begin{array}{l}\int hd\mu_n \to \int hd\mu \\ \text{for } C^2 \text{ test } h\end{array}} \arrow[r, Leftrightarrow] & \boxed{\begin{array}{l}\int \bar{\hat{h}}\varphi_n dt \to \int \bar{\hat{h}}\varphi dt \\ \text{for } C^2 \text{ test } h\end{array}} &
\end{tikzcd}$$

Theorem 3.14 (Characteristic Functions Dictate Distributions)**.** *If random variables X and Y have the same characteristics functions, i.e. $\varphi_X = \varphi_Y$, then X and Y have the same distribution.*

Proof. Consider a sequence $X_n = X$ for all n, and we have $\varphi_{X_n} = \varphi_X = \varphi_Y$ so $X_n \implies Y$. This means, $F_X(t) = F_{X_n}(t) \to F_Y(t)$ at the continuity points of F_Y. For t not a continuity point, we can approximated by continuity points t_n converging to t from above and have the following by right continuity:

$$F_X(t) = \lim_{n\to\infty} F_X(t_n) = \lim_{n\to\infty} F_Y(t_n) = F_Y(t).$$

□

Theorem 3.15 (Properties of Characteristic Functions)**.** *Let X be such that $\mathbb{E}|X|^m < \infty$. Then φ is m times differentiable, $\varphi^{(m)}$ is continuous, and*

$$\varphi^{(k)}(t) = i^k \mathbb{E}[X^k e^{itX}]$$

for $0 \le k \le m$. In particular, if X is square integrable, then we have

1. *φ is twice differentiable and φ'' is continuous at zero.*
2. *$\varphi'(0) = i\mathbb{E}[X]$.*
3. *$\varphi''(0) = -\mathbb{E}[X^2]$.*

Proof. We will prove the first half by induction, and the second half is just a special case of the general result.

When $k = 0$, we know that ϕ is continuous from the properties of a characteristic function, and the equation follows from the definition.

Assume the results hold for some $k < m$, and we will show for $k + 1$. For small h, we have

$$\frac{e^{ihX} - 1}{h} \to (e^{ihX})'|_{h=0} = iX$$

and

$$\left|\frac{e^{ihX} - 1}{h}\right| = \left|\frac{1}{h}\int_0^{hX} e^{is} ds\right| \le |X|.$$

So, we have the following using the dominated convergence theorem,

$$\begin{aligned}\frac{\varphi^{(k)}(t+h) - \varphi^{(k)}(t)}{h} &= i^k \mathbb{E}\left[\frac{X^k e^{i(t+h)X} - X^k e^{itX}}{h}\right] \\ &= i^k \mathbb{E}\left[X^k e^{itX} \frac{e^{ihX} - 1}{h}\right] \to i^{k+1}\mathbb{E}\left[X^{k+1} e^{itX}\right]\end{aligned}$$

which is of the desired form. The continuity can be established using dominated convergence theorem as well by considering any $t_n \to t$. □

3.3 Central Limit Theorem

Theorem 3.16 (Central Limit Theorem)**.** *If $\{X_n\}$ are iid random variables with mean μ and variable σ^2, and $S_n = X_1 + \cdots + X_n$, then we have*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \implies N(0, 1).$$

Proof. WLOG, we can assume X_n have mean 0 and variance 1 since we can always standardise them. We let their characteristic function be φ. Next, we know that the characteristic function of S_n is φ^n and that of $S_n/\sqrt{n}$ is $\varphi^n(t/\sqrt{n})$. As the characteristic function of N(0,1) is $e^{-t^2/2}$, we just need to show $\varphi^n(t/\sqrt{n}) \to e^{-t^2/2}$ for all t to establish the desired weak convergence.

Expand φ around 0 using Taylor series yield

$$\varphi(t) = \varphi(0) + \varphi'(0) + \varphi''(\xi_t)t^2/2 = 1 + \varphi''(\xi_t)t^2/2.$$

Here ξ_t is in between 0 and t. Since φ'' is continuous at 0, we have

$$\varphi''(\xi_t) = \varphi''(0) + \varepsilon(t) = -1 + \varepsilon(t)$$

where $\varepsilon(t) \to 0$ as $t \to 0$. So, we have

$$\varphi(t) = 1 - \frac{t^2}{2} + \frac{\varepsilon(\xi_t)t^2}{2}$$

so

$$\varphi^n(t/\sqrt{n}) = \left(1 - \frac{t^2}{2n} + \frac{\varepsilon''(\xi_t)t^2}{2n}\right)^n \to e^{-t^2/2}$$

as desired. □

Using a similar strategy, we can also obtain the weak law of large numbers.

Theorem 3.17 (WLLN)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean μ. Then, we have*

$$\frac{X_1 + \cdots + X_n}{n} \implies \mu.$$

Proof. The proof is similar to that of the CLT. WLOG, we let μ to be zero. Also, we denote $S_n := X_1 + \cdots + X_n$.

Note that the characteristic function of a random variable that takes 0 with probability one is $e^0 = 1$, so we would like to show $\varphi_{S_n/n}(t) \to 1$ for all t as $n \to \infty$.

Note that from the properties of a characteristic function, we have

$$\varphi_{S_n/n}(t) = \varphi_{S_n}(t/n) = \varphi_X(t/n)^n.$$

Consider $\varphi_X(t)$ and do a Taylor expansion at 0, so we have

$$\varphi_X(t) = \varphi(0) + \varphi'(\xi_t)t = 1 + \varphi'(\xi_t)t$$

for some $\xi_t \in (0, t)$. Next, we know that φ' is continuous at 0 so

$$\varphi'(\xi_t) = \varphi'(0) + \varepsilon(t) = \varepsilon(t) \to 0$$

as $t \to 0$. So,

$$\varphi_{S_n/n}(t) = \varphi_X(t/n)^n = [1 + \varepsilon(t/n)t/n]^n \to 1^n = 1$$

as $n \to \infty$. Done. □

Chapter 4

Martingales

4.1 Conditional Expectation

Theorem 4.1. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, X Be an integrable random variable, and $F \subset \Sigma$ be a sub-σ-algebra. Then, there exists a random variable Y such that*

1. *Y is F-measurable.*
2. *Y is integrable.*
3. *$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}$ for all $A \in F$.*

Y is unique almost surely, i.e. if $\tilde{Y}$ satisfies (1) - (3) then $Y = \tilde{Y}$ a.s. In particular, Y is denoted by $\mathbb{E}[X \mid F]$, and is called the ***conditional expectation*** *of X given F.*

The proof of this theorem requires absolutely continuous measures and the Radon-Nikodym theorem. Consider measures Q, P, we say Q is **absolutely continuous** w.r.t. P, denoted by $Q << P$, if $P(A) = 0 \implies Q(A) = 0$. The Radon-Nikodym theorem states that, if P and Q are σ-finite measures, and $Q << P$, then there exists a measurable integrable Y such that $Q(A) = \int_A Y dP$. In particular, Y, sometimes denoted by dQ/dP, is called the **Radon-Nikodym derivative** of Q w.r.t. P.

Proof. WLOG, we assume $X \geq 0$. The case of general X can be easily extended from this case, as we have $X = X_+ - X_-$.

Consider the measure space (Ω, F), and 2 measures on it: (1) probability measure $\mathbb{P}$ restricted to F, and (2) Q with $Q(A) = \int_A X d\mathbb{P}$. $\mathbb{P}$ is finite as it is a probability measure, and Q is finite since X is integrable. Next, if we have $\mathbb{P}(A) = 0$, then $Q(A) = \int_A X d\mathbb{P} = 0$. So, $Q << P$. Therefore, using the Radon-Nikodym theorem, there exists a measurable, integrable Y such that Q has density Y w.r.t. $\mathbb{P}$.

Then, $Q(A) = \int_A X d\mathbb{P} = \int_A Y d\mathbb{P}$ for all $A \in F$, so we have

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}$$

for all $A \in F$, as desired. This has demonstrated the existence.

Suppose there exists $\tilde{Y}$ satisfying (1) - (3) and $\mathbb{P}(\tilde{Y} \neq Y) > 0$. Notice that $\{\tilde{Y} \neq Y\} = \cup_{n=1}^{\infty}\{Y - \tilde{Y} > 1/n\} \cup \cup_{n=1}^{\infty}\{Y - \tilde{Y} < -1/n\}$. So, we have

$$0 = \int_A Y d\mathbb{P} - \int_A \tilde{Y} d\mathbb{P} = \int_A Y - \tilde{Y} d\mathbb{P} \geq \frac{1}{n}\mathbb{P}(A) > 0$$

where $A = \{Y - \tilde{Y} > 1/n\}$ for some n. This is a contradiction. Thus, we have established uniqueness too. □

Theorem 4.2 (Properties of Conditional Expectation)**.** *We have*

1. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|F]]$.
2. *If X is F-measurable, then* $\mathbb{E}[X|F] = X$.
3. *(Linearity)* $\mathbb{E}[a_1X_1 + a_2X_2 \mid F] = a_1\mathbb{E}[X_1|F] + a_2\mathbb{E}[X_2|F]$.
4. *(Positivity) If $X \geq 0$ a.s., then $\mathbb{E}[X|F] \geq 0$ for all $F \subset \Sigma$.*
5. *(Conditional Monotone Convergence Theorem) If $0 \leq X_n \nearrow X$, then $0 \leq \mathbb{E}[X_n|F] \nearrow \mathbb{E}[X|F]$.*
6. *(Taking Out What's Known) If Z is F-measurable, then $\mathbb{E}[ZX|F] = Z\mathbb{E}[X|F]$.*
7. *(Independence) If Z is independent of F, then $\mathbb{E}[Z|F] = \mathbb{E}[Z]$.*
8. *(Tower Property) Consider σ-algebras $G \subset F \subset \Sigma$, then $\mathbb{E}[\mathbb{E}[X|F]|G] = \mathbb{E}[X|G]$.*
9. *(Conditional Jensen) For convex φ, $\mathbb{E}[\varphi(X)|F] \geq \varphi(\mathbb{E}[X|F])$.*

Example. Consider a sequence of iid $\{X_n\}$ with mean μ. We have

$$\begin{aligned}
&\mathbb{E}[X_1 + \cdots + X_n \mid \sigma(X_1, \ldots, X_m)] \\
&= \mathbb{E}[X_1 + \cdots + X_m \mid \sigma(X_1, \ldots, X_m)] + \mathbb{E}[X_{m+1} + \cdots + X_n \mid \sigma(X_1, \ldots, X_m)] \\
&= X_1 + \cdots + X_m + \mu(n - m).
\end{aligned}$$

Example. Consider a random variable X that takes p_1 with probability q and takes p_2 with probability $1 - q$. Both p_1 and p_2 are in between 0 and 1. Next, we have a random variable Y that is a Bernoulli random variable with parameter X. So, we have $Y = Y_1 1_{\{X=p_1\}} + Y_2 1_{\{X=p_2\}}$, where Y_1 is Bernoulli(p_1) and Y_2 is Bernoulli(p_2). We have

$$\begin{aligned}
&\mathbb{E}[Y|\sigma(X)] \\
&= \mathbb{E}[Y_1 1_{\{X=p_1\}} \mid \sigma(X)] + \mathbb{E}[Y_2 1_{\{X=p_2\}} \mid \sigma(X)] \\
&= 1_{\{X=p_1\}}\mathbb{E}[Y_1] + 1_{\{X=p_2\}}\mathbb{E}[Y_2] \\
&= p_1 1_{\{X=p_1\}} + p_2 1_{\{X=p_2\}} \\
&= X.
\end{aligned}$$

4.2 Martingales

Definition 4.3. *A **filtration** on $(\Omega, \Sigma, \mathbb{P})$ is a growing sequence of σ-algebras $F_0 \subseteq F_1 \subseteq \cdots \subset \Sigma$. If $\{X_n\}$ is a stochastic process, then $\sigma(X_0) \subseteq \sigma(X_0, X_1) \subseteq \cdots \subset \Sigma$ is called the **natural filtration** of $\{X_n\}$.*

*A stochastic process $\{X_n\}$ is said to be **adapted** to filtration $\{F_n\}$ if X_n is F_n-measurable for all n.*

Definition 4.4. *$\{X_n\}$ is a **martingale** with regards to the filtration $\{F_n\}$ if we have*

1. *$\{X_n\}$ is adapted to F_n.*

2. $\mathbb{E}[|X_n|] < \infty$ *for all* n.
3. $\mathbb{E}[X_{n+1}|F_n] = X_n$ *for all* n.

A ***submartingale*** *is a martingale but with* $\mathbb{E}[X_{n+1}|F_n] \geq X_n$ *instead. A* ***supermartingale*** *is a martingale but with* $\mathbb{E}[X_{n+1}|F_n] \leq X_n$ *instead.*

Remark. A submartingale can be transformed into a supermartingale (and vice versa) by adding a minus sign to each random variable. As a result, we would only consider either submartingales or supermartingales.

Remark. A martingale (in real life) is a tool to make sure a horse is staying in its lane properly. Here, the random variables of a martingale stay the same in expectation at each step, which is why it is given this name.

Proposition 4.5. *For a martingale* $\{X_n\}$*, we have*

1. $\mathbb{E}[X_n|F_n] = X_m$ *for all* $m < n$.
2. $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ *for all* n.

Proof. The first result relies on the tower property of conditional expectations, while the second result relies on the iterated expectation. □

Example. Consider $\{X_n\}$ with

$$X_n = \sum_{i=1}^{n} Y_i \varepsilon_i$$

with filtration $\{F_n\}$ and $F_n = \sigma(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n)$. Here, $\{\varepsilon_i\}$ is an iid Bernoulli(1/2) sequence and $\{Y_i\}$ is a random sequence with each Y_n being F_{n-1} measurable.

In this case, we have

$$\begin{aligned}\mathbb{E}[X_{n+1}|F_n] &= \sum_{i=1}^{n} Y_i \varepsilon_i + \mathbb{E}[Y_{n+1}\varepsilon_{n+1}|F_n] \\ &= X_n + Y_{n+1}\mathbb{E}[\varepsilon_{n+1}] \\ &= X_n.\end{aligned}$$

4.3 Stopping Time

Definition 4.6. $T : \Omega \to \mathbb{N} \cup \{+\infty\}$ *is called a* ***stopping time*** *if* $\{T = n\} \in F_n$ *for all* $n \in \mathbb{N}$*. Here* $\{F_n\}$ *is a filtration.*

Proposition 4.7. *For stopping times* S, T *with respect to the same filtration* (F_n)*,*

1. $S + T$ *is a stopping time.*
2. $S - T$ *is* ***not*** *a stopping time.*
3. ST *is a stopping time.*

Proof. A stopping time K must satisfy $\{K = k\} \in F_k$ for all k.

We notice that

$$\{S + T = k\} = \bigcup_{q=0}^{k} \{S = q, T = k - q\} = \bigcup_{q=0}^{k} \left[\{S = q\} \cap \{T = k - q\}\right] \in F_k$$

since both $\{S = q\}$ and $\{T = k - q\}$ are in F_k. Thus, $S + T$ is a stopping time.

Using the same logic, we can notice that $S - T$ is not a stopping time. Say $T = 1$, and $S - T = k \iff S = k + 1$ which is not in F_k.

Similarly, we have

$$\{ST = k\} = \bigcup_q \{S = q, T = k/q\} = \bigcup_q [\{S = q\} \cap \{T = k/q\}] \in F_k$$

since both $\{S = q\}$ and $\{T = k/q\}$ are in F_k. □

For a martingale $\{X_n\}$ with filtration $\{F_n\}$ and stopping time T, X_T is known as the stopped random variable, while $\{X_{n\wedge T}\}$ is a stopped process. A stopped random process will behave like the original martingale until the stopping time, and stay at X_T from T onwards.

Theorem 4.8. *If $\{X_n\}$ is a martingale with regards to filtration $\{F_n\}$, and T is a stopping time then the stopped process $\{X_{n\wedge T}\}$ is a martingale as well.*

Theorem 4.9 (Optimal Stopping Theorem, OST)**.** *If $\{X_n\}$ is a martingale with regards to filtration $\{F_n\}$ and T is a stopping time that is finite a.s., then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$ in each of the following 3 situations*

1. *T is bounded a.s.*
2. *$\mathbb{E}[T] < \infty$ and (X_n) has a.s. bounded increments.*
3. *$\{X_n\}$ is bounded a.s.*

Proof. We have $X_{n\wedge T} \to X_T$ a.s. as $n \to \infty$, and if we can satisfy the condition of the dominated convergence theorem, then we would have

$$\mathbb{E}[X_{n\wedge T}] \to \mathbb{E}[X_T].$$

Since $(X_{n\wedge T})$ is a martingale, we have $\mathbb{E}[X_{n\wedge T}] = \mathbb{E}[X_0]$ which is a constant, so we would have the desired $\mathbb{E}[X_T] = \mathbb{E}[X_0]$ as a result of the convergence. So, we just need to make sure in each of the three situations we have satisfied the conditions of the dominated convergence theorem.

(2) Notice that $\{X_n\}$ has bounded increments, so let the increment be bounded by c. Then, we have

$$|X_{n\wedge T} - X_0| \leq \sum_{i=0}^{n\wedge T} |X_{i+1} - X_i| \leq c(n \wedge T + 1) \leq c(T + 1)$$

and we have $\mathbb{E}[T] < \infty$. This satisfied the desired conditions.

(3) Since $\{X_n\}$ is bounded, we have $|X_{n\wedge T}| \leq |X_n|$ which is bounded as well. Done.

(1) We would not use dominated convergence to establish the desired result in this case. Since we have $T \leq N$ for some N, for all $n \geq N$, we have $X_{n\wedge T} = X_T$, so $\mathbb{E}[X_{n\wedge T}] = \mathbb{E}[X_T]$, which establish the desired convergence. □

Remark. It is not hard to notice that the proof would still work if we have $\{X_{n\wedge T}\}$ instead of $\{X_n\}$ for situations 2 and 3. This in fact gives a stronger result.

4.4 Strong Law of Large Number

Theorem 4.10 (Doob's Submartingale Inequality). *Let $\{X_n\}$ be a non-negative submartingale. Then $c\mathbb{P}(\max_{0\leq n\leq N} X_n \geq c) \leq \mathbb{E}[X_N]$.*

Proof. Let $T := \inf\{n : X_n \geq c\} \wedge N$. Clearly, $T \leq N$. If we have $T \leq N \implies \mathbb{E}(X_T) \leq \mathbb{E}(X_N)$, then we would also have $\mathbb{E}(X_T 1_E) \leq \mathbb{E}(X_N 1_E) \leq \mathbb{E}(X_N)$ where $E = \{\max_{0\leq n\leq N} X_n \geq c\}$ as

$$\mathbb{E}(X_T) = \mathbb{E}(X_T 1_E) + \mathbb{E}(X_T 1_{E^c}) \leq \mathbb{E}(X_N 1_E) + \mathbb{E}(X_N 1_{E^c}) = \mathbb{E}(X_N).$$

This gives us $c\mathbb{P}(E) \leq \mathbb{E}(X_T 1_E) \leq \mathbb{E}(X_N)$ as required.

Now we just need to show $T \leq N \implies \mathbb{E}(X_T) \leq \mathbb{E}(X_N)$.

We have

$$\begin{aligned}
\mathbb{E}(X_T) &= \mathbb{E}\left[\sum_{i=0}^N X_T 1_{T=i}\right] = \sum_{i=0}^N \mathbb{E}\left[X_T 1_{T=i}\right] = \sum_{i=0}^N \mathbb{E}\left[X_i 1_{T=i}\right] \\
&= \sum_{i=0}^N \int_{T=i} X_i d\mathbb{P} \leq \sum_{i=0}^N \int_{T=i} \mathbb{E}(X_N|F_i) d\mathbb{P} = \sum_{i=0}^N \int_{T=i} X_N d\mathbb{P} \\
&= \sum_{i=0}^N \mathbb{E}[X_N 1_{T=i}] = \mathbb{E}[X_N]\sum_{i=0}^N \mathbb{E}[1_{T=i}] = \mathbb{E}[X_N]\sum_{i=0}^N \mathbb{P}[T=i] = \mathbb{E}[X_N].
\end{aligned}$$

□

Lemma 4.11 (Kolmogorov Inequality). *Let $\{X_n\}$ be independent, square integrable, mean 0 random variables. Then*

$$c^2\mathbb{P}\left(\max_{1\leq n\leq N}\left|\sum_{i=1}^n X_i\right| \geq c\right) \leq \sum_{i=1}^N \mathbb{E}[X_i^2].$$

This is a direct consequence of the Doob's submartingale inequality.

Theorem 4.12 (Kolmogorov Theorem). *Let $\{X_n\}$ be independent, square integrable, mean 0 random variables. If $\sum_{n=1}^\infty \mathbb{E}(X_n^2) < \infty$, then $\sum_{n=1}^\infty X_n < \infty$ a.s.*

Before proving this result, let us first look at an example.

Example. We know that $\sum 1/n = \infty$ and $\sum(-1)^n/n < \infty$ from Analysis. If we have $\sum \varepsilon_n/n$ with ε_n be i.i.d. random variables taking ±1 with equal probability, then we know that

$$\mathbb{E}\left[\frac{\varepsilon_n}{n}\right] = 0, \quad \sum \mathbb{E}\left[\frac{\varepsilon_n^2}{n^2}\right] = \sum \frac{1}{n^2} < \infty,$$

so we know $\sum \varepsilon_n/n < \infty$ by Kolmogorov Theorem above. Furthermore, by staring at the derivation above, we have $\sum \varepsilon_n/n^\alpha < \infty$ for all $\alpha > 1/2$.

Now, let us prove the theorem.

Proof. Notice that $\sum_{n=1}^{\infty} X_n < \infty$ is equivalent to say that the sequence of partial sums converges, and thus the sequence is a Cauchy sequence. So, by the definition of a Cauchy sequence, we would like to show

$$\mathbb{P}\left(\forall\varepsilon, \exists N\ s.t. \forall n > N, \left|\sum_{i=N+1}^{n} X_i\right| < \varepsilon\right) = 1.$$

This is equivalent as

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k=1}^{\infty}\bigcup_{N=1}^{\infty}\bigcap_{n>N}\left\{\left|\sum_{i=N+1}^{n} X_i\right| < \frac{1}{k}\right\}\right) &= 1 \\
\mathbb{P}\left(\bigcup_{N=1}^{\infty}\bigcap_{n>N}\left\{\left|\sum_{i=N+1}^{n} X_i\right| < \frac{1}{k}\right\}\right) &= 1 \quad \text{for each } k \\
\mathbb{P}\left(\bigcup_{N=1}^{\infty}\bigcup_{n=N+1}^{\infty}\left\{\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right\}\right) &= 0 \quad \text{for each } k.
\end{aligned}$$

Now, using the Kolmogorov inequality, we have

$$\mathbb{P}\left(\bigcup_{n=N+1}^{M}\left\{\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right\}\right) = \mathbb{P}\left(\max_{N+1\leq n\leq M}\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right) \leq k^2 \sum_{n=N+1}^{M} \mathbb{E}[X_n^2].$$

Taking the limit of $M \to \infty$ gives us

$$\mathbb{P}\left(\bigcup_{n=N+1}^{\infty}\left\{\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right\}\right) \leq k^2 \sum_{n=N+1}^{\infty} \mathbb{E}[X_n^2].$$

So, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{N=1}^{\infty}\bigcup_{n=N+1}^{\infty}\left\{\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right\}\right) &\leq \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{n=N+1}^{\infty}\left\{\left|\sum_{i=N+1}^{n} X_i\right| \geq \frac{1}{k}\right\}\right) \\
&\leq \lim_{N\to\infty} k^2 \sum_{n=N+1}^{\infty} \mathbb{E}[X_n^2] = 0
\end{aligned}$$

since $\sum_{n=1}^{\infty} \mathbb{E}[X_n^2] < \infty$. □

Next, we can prove the strong law of large number for square integrable random variables. But first, let us have some auxiliary lemmas.

Lemma 4.13 (Cesaro). *If $a_n \to a$, then $(a_1 + \cdots + a_n)/n \to a$.*

Proof. As we have $a_n \to a$, for $\varepsilon > 0$, we have N such that for all $n > N$, $|a_n - a| < \varepsilon$. Then, we have

$$\left|\frac{1}{n}(a_1 + \cdots + a_n) - a\right| \leq \left|\frac{a_1 + \cdots + a_M}{n} - \frac{M}{n}a\right| + \frac{n-M}{n}\varepsilon < \varepsilon$$

for sufficiently large M and $n > M$. □

Lemma 4.14 (Kronecker). *If $\sum_{n=1}^{\infty} a_n/n < \infty$, then $(a_1 + \cdots + a_n)/n \to 0$.*

Proof. Denote $u_0 = 0$, $u_n = \sum_{i=1}^{n} a_i/i$ and $u = \sum_{i=1}^{\infty} a_i/i$. We have $u_n \to u$, and $a_n = n(u_n - u_{n-1})$. So,

$$\begin{aligned}\frac{a_1 + \cdot + a_n}{n} &= \frac{1}{n}(u_1 - u_2 + 2(u_2 - u_3) + \cdots + n(u_n - u_{n-1})) \\ &= \frac{1}{n}(nu_n - u_0 - u_1 - \cdots - u_{n-1}) \\ &= u_n - \frac{u_0 + u_1 + \cdots + u_{n-1}}{n} \to u - u = 0.\end{aligned}$$

□

Let us prove the SLLN for square integrable random variables.

Theorem 4.15 (SLLN for square integrable random variables)**.** *Let $\{X_i\}$ be i.i.d. square integrable random variables with mean 0. Then,*

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to 0 \qquad a.s.$$

Proof. Let $Y_n = X_n n$. Then,

$$\sum_{n=1}^{\infty} \mathbb{E}[Y_n^2] = \mathbb{E}[X_1^2] \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Using Kolmogorov's Theorem, we have $\sum_{n=1}^{\infty} Y_n < \infty$ a.s., so $\sum_{n=1}^{\infty} X_n/n < \infty$ a.s.. Then we can use the Kronecker lemma to complete the proof. □

If we would like to remove the square integrability condition, we need to have a modified sequence that is close enough to the original sequence and consists of square integrable random variables. We would need another result in order to do so.

Theorem 4.16 (Kolmogorov Truncation Lemma (TL))**.** *Let $\{X_n\}$ be i.i.d. integrable random variables with $\mathbb{E}[X_n] = \mu$. Define the truncated random variables $Y_n = X_n 1_{\{|X_n| \leq n\}}$. Then,*

1. *$\mathbb{P}(X_n = Y_n$ eventually$) = \mathbb{P}(\exists N\ \forall n \geq N\ X_n = Y_n) = 1$.*
2. *$\mathbb{E}[Y_n] \to \mu$.*
3. *$\sum_{n=1}^{\infty}$ Var$(Y_n)/n^2 < \infty$.*

Proof. (1) We want $\mathbb{P}(X_n \neq Y_n\ i.o.) = 0$, which is $\mathbb{P}(|X_n| > n\ i.o.) = 0$. This follows from Borel-Cantelli 1 as we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) = \sum_{n=1}^{\infty} \mathbb{E}[1_{\{|X_1|>n\}}] = \mathbb{E}\left[\sum_{n=1}^{\infty} 1_{\{|X_1|>n\}}\right] \leq \mathbb{E}[|X_1|] < \infty$$

where the last equality is because of the monotone convergence theorem.

(2) We have

$$\mathbb{E}[Y_n] = \mathbb{E}[X_n 1_{\{|X_n| \leq n\}}] = \mathbb{E}[X_1 1_{\{|X_1| \leq n\}}] \to \mathbb{E}[X_1] = \mu.$$

where the convergence is due to dominated convergence theorem.

(3) We have

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}[Y_n^2]}{n^2} = \sum_{n=1}^{\infty} \frac{\mathbb{E}[X_1^2 1_{\{|X_1|\le n\}}]}{n^2} = \mathbb{E}\left[\sum_{n=1}^{\infty} \frac{X_1^2 1_{\{|X_1|\le n\}}}{n^2}\right] = \mathbb{E}\left[X_1^2 \sum_{n\ge |X_1|}^{\infty} \frac{1}{n^2}\right].$$

Note that we have

$$\sum_{n\ge m}^{\infty} \frac{1}{n^2} \le \sum_{n\ge m}^{\infty} \frac{2}{n(n+1)} = \frac{2}{m}$$

for integer m. Therefore, we will ignore the fact that $|X_1|$ does not have to be an integer (it is only technicality for the cases when it is not) and assume that it is, and we have

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}[Y_n^2]}{n^2} \le \mathbb{E}\left[X_1^2 \frac{2}{|X_1|}\right] = 2\mathbb{E}[|X_1|] < \infty.$$

Since we know $\mathbb{E}[Y_n]$ converges, $\mathbb{E}[Y_n]^2$ is bounded by some c and

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}[Y_n]^2}{n^2} \le \sum_{n=1}^{\infty} \frac{c}{n^2} < \infty.$$

So we have the desired convergence. □

Now, let us prove the SLLN.

Theorem 4.17 (SLLN). *Let $\{X_n\}$ be i.i.d. integrable random variables with $\mathbb{E}[X_i] = \mu$. Then,*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \qquad a.s.$$

Proof. We first defined Y_n as in the case of the truncated lemma. Then, we just need to show that

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \to \mu \qquad \text{a.s.}$$

by TL(1). Since TL(2) says $\mathbb{E}[Y_n] \to \mu$, we have, using Cesaro,

$$\frac{\mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_n]}{n} \to \mu$$

so we just need to show that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) \to 0 \qquad \text{a.s.}$$

Notice that $Y_i - \mathbb{E}[Y_i]$ are independent with mean 0, and

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2]}{i^2} = \sum_{i=1}^{\infty} \frac{\mathrm{Var}[Y_i]}{i^2} < \infty$$

by TL(3). So, the desired result follows from Kolmogorov theorem and Kronecker lemma. □

4.5 Martingale Convergence Theorem*

Theorem 4.18 (Martingale Convergence Theorem)**.** *Let $\{X_n\}$ be a martingale bounded in L_1. Then, there exists a random variable X defined on the same probability space such that $X_n \to X$ a.s.*

Corollary 4.19. *A non-negative martingale converges to a random variable a.s.*

Proof. For a non-negative martingale $\{X_n\}$, we know that $\mathbb{E}[|X_n| = \mathbb{E}[X_n] = \mathbb{E}[X_0]$ where the last equality follows from $\mathbb{E}[X_n|F_n] = X_0$ and $\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n|F_n]] = \mathbb{E}[X_0]$. So, every non-negative martingale is bounded in L_1 and thus converges a.s. by the martingale convergence theorem. □

4.6 Galton-Watson Process

In this section, we will investigate various interesting properties of the Galton-Watson process.

The Galton-Watson process was proposed to study the extinction of surnames. The same process can be used to model other things, such as the survival of progeny of mutant genes. In this case, we will think about this model in its original surname context.

Let $\{Z_n\}_{n\in\mathbb{N}}$ denote the stochastic process of number of males with a particular surname, and we let $Z_0 := 1$. So, Z_n denotes the number of males with a particular surname in the n-th generation. It is natural to assume that these random variables take values in $\mathbb{N}$. To study the process requires us to know how many (male) offspring each male has. We **assume** that each male will have X offspring, where X is a random variable take takes values in $\mathbb{N}$ and $\mathbb{P}(X = 0) > 0$. Note that we also assume that the number of offspring for each male is independent.

To tidy up the notations, we denote X_r^{m+1} as the number of offspring of the r-th male of generation m. Clearly, r takes values between 1 and Z_m. Consequently, we have

$$Z_{m+1} = \sum_{r=1}^{Z_m} X_r^{m+1}.$$

It is obvious that $\{Z_n\}$ is a Markov chain where the Markov property follows from the independence of the number of offspring for each male. The Markov property states that condition on the current state, the distribution of the future states is independent of past states.

Let $\sigma(Z_n)$ to denote the σ-algebra of the random variable Z_n. We would like to know $\mathbb{E}[Z_{n+1}|\sigma(Z_n)]$. We have

$$\begin{aligned}\mathbb{E}[Z_{n+1}|\sigma(Z_n)] &= \mathbb{E}\left[\sum_{i=0}^{\infty} Z_{n+1}1_{\{Z_n=i\}}|\sigma(Z_n)\right] = \sum_{i=0}^{\infty}\mathbb{E}\left[\sum_{r=1}^{i} X_r^n 1_{\{Z_n=i\}}|\sigma(Z_n)\right] \\ &= \sum_{i=0}^{\infty} 1_{\{Z_n=i\}}\mathbb{E}\left[\sum_{r=1}^{i} X_r^n\right] = \sum_{i=0}^{\infty} 1_{\{Z_n=i\}} i\mathbb{E}\left[X\right] = aZ_n,\end{aligned}$$

where $a := \mathbb{E}[X]$.

Notice that when $a = 1$, $\{Z_n\}$ is a martingale. And when $a > 1$ and $a < 1$ the process is a submartingale and a supermartingale respectively.

Of course, if we consider the process $\{Z_n/a^n\}$ instead, then we would get a martingale all the time.

Next, we would want to know how would Z_n behave as $n \to \infty$. This certainly depends on the value of a. When $a > 1$, the process will grow exponentially fast (with probability 1). When $a < 1$, it could be shown that the process will die out within finite number of generations. The problem is interesting when $a = 1$.

There is a boring scenario for $a = 1$, which is when $\mathbb{P}(X = 1) = 1$. In this case, $Z_n = 1$ for all n. Things get interesting when $\mathbb{P}(X = 1) < 1$.

In this case, martingale convergence theorem tells us that $Z_n \to Z$ a.s. for some random variable Z. Since each Z_n is an integer, Z is also taking integer values. We would like to know which integer k would Z take with positive probability.

For each integer $k \neq 0$, consider

$$\{Z_n \to k\} = \{\exists N, \forall n \geq N, Z_n = k\} = \bigcup_{N=1}^{\infty} \{n \geq N, Z_n = k\}.$$

Let $q(k)$ denotes the probability that k males producing k offspring, and this probability is strictly less than 1. Then,

$$\mathbb{P}\{n \geq N, Z_n = k\} = \mathbb{P}(Z_N = k) \lim_{M \to \infty} q(k)^{M-N} = 0$$

so $\mathbb{P}\{Z_n \to k\} = 0$. Thus, $Z_n \to 0$, and the process dies out in this case too.

Appendix A

Measure Theory Basics

Definition A.1. *For a set S, a collection Σ of subsets of S is called an **σ-algebra** on S if*

1. $\emptyset, S \in \Sigma$
2. $F \in \Sigma \implies F^C = S \backslash F \in \Sigma$
3. $\{F_n\}_{n\in\mathbb{N}} \subset \Sigma \implies \cup_n F_n \in \Sigma$.

Note that the closure under countable union condition contains closures under finite union as we can take the rest of the sequence to be empty set. If we only have closure under finite union, the collection is an **algebra**.

A set X with an σ-algebra $\mathcal{A}$ defined on it form a pair $(X, \mathcal{A})$ that is known as a **measurable space**.

Definition A.2. *Given a measurable space $(X, \mathcal{A})$, a function $\mu : \mathcal{A} \to [0, \infty]$ is a **measure** if we have*

1. $\mu(\emptyset) = 0$
2. *For a sequence of disjoint sets $\{E_j\}_{j=1}^\infty \subset \mathcal{A}$, we have $\mu(\cup_{j=1}^\infty E_j) = \sum_{j=1}^\infty \mu(E_j)$.*

A measure is **finite** if we also have $\mu(X) < \infty$, and a measure is **σ-finite** if we also have $\mu(E_j) < \infty$ for all j for $E_j \in \mathcal{A}$ and $X = \cup_{j=1}^\infty E_j$.

A.1 Construction of Measure

In this section we look at how one could construct a measure abstractly. There is a four-step procedure.

1. Take an semi-ring $\mathcal{A}$ on set X
2. Define a premeasure μ_0 on $\mathcal{A}$
3. Construct an outer measure μ^* from premeasure
4. Obtain a measure by restricting the outer measure to outer measurable sets

In the following we will explain in detail this procedure.

Step 1

Definition A.3. *A collection $S \subset \mathcal{P}(X)$*[1] *is a* ***semi-ring*** *on X if*

1. $\emptyset \in S$
2. $A, B \in S \implies A \cap B \in S$
3. *If $A, B \in S$, then there exists finitely many disjoints $D_1, \ldots, D_N \in S$ such that $A \backslash B = \cup_{j=1}^{N} D_j$.*

Step 2

Definition A.4. *Let S be a semi-ring. A function $\mu_0 : S \to [0, \infty]$ is called a* ***premeasure*** *if it satisfies*

1. $\mu_0(\emptyset) = 0$
2. *Whenever $\{R_j\}_{j=1}^{\infty} \subset S$ are pairwise disjoint and satisfy $\cup R_j \in S$, then $\mu_0(\cup R_j) = \sum \mu_0(R_j)$.*

Step 3

Definition A.5. *For a nonempty set X, a function $\mu^* : X \to [0, \infty]$ is an* ***outer measure*** *if it satisfies*

1. $\mu^*(\emptyset) = 0$
2. $A \subset B \implies \mu^*(A) \leq \mu^*(B)$
3. $\mu^*(\cup_{j=1}^{\infty} A_j) \leq \sum_{j=1}^{\infty} \mu^*(A_j)$.

We can construct an outer measure in the following way.

Proposition A.6. *Let $\xi \subset \mathcal{P}(X)$ with $\emptyset, X \in \xi$. Let $\rho : \xi \to [0, \infty]$ be such that $\rho(\emptyset) = 0$. Then, there exists an outer measure μ^* on X defined by,*

$$\mu^*(A) = \inf \left\{ \sum_{j=1}^{\infty} \rho(E_j) \mid E_j \in \xi, A \subset \bigcup_{j=1}^{\infty} E_j \right\}$$

for all $A \subset X$.

Remark. We notice that for the definition to work, we need ξ to contain a cover for every subset A. We would call ξ as the **covering class**.

Proof. It is easy to see by definition that $\mu^*(\emptyset) = 0$ as we can have $E_j = \emptyset$.

For $A \subset B$, we know that for every sequence of E_j such that $B \subset \cup E_j$, we have $A \subset B \subset \cup E_j$, so we are taking the infimum over a larger set for $\mu^*(B)$ than for $\mu^*(A)$.

Finally, fix some $\varepsilon > 0$. For each A_j, let us have a sequence $\{E_k^j\}_k$ in ξ such that $A_j \subset \cup_k E_k^j$ and $\sum_k \rho(E_k^j) \leq \mu^*(A_j) + \varepsilon/2^j$. Then, as $\cup_j A_j \subset \cup_j \cup_k E_k^j$, we have

$$\mu^*(\cup_j A_j) \leq \sum_j \sum_k \rho^*(E_k^j) \leq \sum_j [\mu^*(A_j) + \varepsilon/2^j] = \sum_j \mu^*(A_j) + \varepsilon.$$

Taking ε to zero gives us the desired result.

□

[1] $\mathcal{P}(X)$ is the power set of X.

Notice that here we do not impose much condition on S and ρ. One of the goal of this result is to extend the notion of distance through ρ, but the outer measure we obtain might not satisfy this goal as we may not have μ^* and ρ agreeing on all $E \in \xi$. Consider this example.

Example. $X = \mathbb{R}$ and $\xi = \{\emptyset, [0,2], [2,4], [1,3], \mathbb{R}\}$. Let us assign the elements of ξ as follows:

$$\rho(\emptyset) = 0, \quad \rho([0,2]) = 3, \quad \rho([2,4]) = 1, \quad \rho([1,3]) = 5, \quad \rho(\mathbb{R}) = \infty.$$

Then, we have $\mu^*([1,3]) \leq 1 + 3 = 4$ as $[1,3] \subset [0,2] \cup [2,4]$ yet $\rho([1,3]) = 5$.

So, more conditions need to be imposed and we have the following result. Before we do that, we need the notion of a μ^*-measurable set.

Definition A.7. *For a nonempty set X, suppose we have an outer measure μ^* on X. We say that a set $A \subset X$ is μ^***-measurable**, or simply **measurable**, if*

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^C)$$

for every $E \subset X$.

Now we are ready to state our stronger result.

Proposition A.8. *Let $S \subset \mathcal{P}(X)$ be a semi-ring with $X \in S$. Suppose that $\rho : S \to [0, \infty]$ is a premeasure. The outer measure we constructed as in the case for Proposition A.6 agree with ρ on S and every element of S is μ^*-measurable.*

Step 4

We are ready to obtain a measure.

Theorem A.9 (Caratheodory Theorem)**.** *Let X be nonempty and μ^* be an outer measure on it. Then,*

1. *The collection $\mathcal{A}$ of μ^*-measurable sets is an σ-algebra.*
2. *The restriction of μ^* to A is a complete measure[2], which we denote it by μ.*

Proof. (1) We will show that $\mathcal{A}$ is an σ-algebra by showing it is closed under complement and closed under countable unions. Closed under complement is easy to see as if A is countable, we have

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^C) = \mu^*(E \cap A^C) + \mu^*(E \cap (A^C)^C) = \mu^*(E)$$

for any subset E.

To show it is closed under countable unions, we first show that it is closed under finite unions, then extend it to the countable case.

For $A, B \in \mathcal{A}$, we have

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap A) + \mu^*(E \cap A^C) \\ &= \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^C) + \mu^*(E \cap A^C \cap B) + \mu^*(E \cap A^C \cap B^C) \\ &= \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^C) + \mu^*(E \cap A^C \cap B) + \mu^*(E \cap (A \cup B)^C).\end{aligned}$$

[2] A measure is complete if it contains every subset of sets with measure zero.

Notice that we have

$$A \cup B = (A \cap B) \cup (A^C \cap B) \cup (A \cap B^C)$$
$$E \cap A \cup B = (E \cap A \cap B) \cup (E \cap A^C \cap B) \cup (E \cap A \cap B^C)$$

so we have

$$\mu^*(E \cap A \cup B) \leq \mu^*(E \cap A \cap B) + \mu^*(E \cap A^C \cap B) + \mu^*(E \cap A \cap B^C),$$

which implies

$$\mu^*(E) \geq \mu^*(E \cap (A \cup B)) + \mu^*(E \cap (A \cup B)^C).$$

The reverse inequality is trivial as $E = [E \cap (A \cup B)] \bigcup [E \cap (A \cup B)^C]$. The finite intersection follows from this via induction.

Next, we will extend it to the countable case. We just need to show it is closed under countable disjoint union, as the rest will follow quite trivially (just remove the union of all the previous sets from the n-th set, which turn an arbitrary sequence to a disjoint sequence).

Consider the sequence of disjoint sets $\{A_j\}$. We define

$$B_n := \bigcup_{j=1}^{n} A_j, \quad B := \bigcup_{j=1}^{\infty} A_j$$

and we have

$$B_n \cap A_n = A_n, \quad B_n \cap A_n^C = B_{n-1}.$$

So, considering arbitrary subset E, we have

$$\mu^*(E \cap B_n) = \mu^*(E \cap B_n \cap A_n) + \mu^*(E \cap B_n \cap A_n^C) = \mu^*(E \cap A_n) + \mu^*(E \cap B_{n-1}).$$

This then implies $\mu^*(E \cap B_n) = \sum_{j=1}^{n} \mu^*(E \cap A_j)$. Using this result, we have

$$\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^C) = \sum_{j=1}^{n} \mu^*(E \cap A_j) + \mu^*(E \cap B_n^C) \geq \sum_{j=1}^{n} \mu^*(E \cap A_j) + \mu^*(E \cap B^C)$$

as $B^C \subset B_n^C$. Taking $\lim_n$ gives us

$$\mu^*(E) \geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \cap B^C) \geq \mu^*(\cup_{j=1}^{\infty} E \cap A_j) + \mu^*(E \cap B^C) \geq \mu^*(E \cap B) + \mu^*(E \cap B^C)$$

as $E \cap B \subset \cup_{j=1}^{\infty} E \cap A_j$. The reverse inequality is trivial, and thus we have obtained closure under countable union.

(2) We need to show that the restriction is a complete measure. First we show the restriction is a measure, then we show it is complete.

From before, notice that we have obtained

$$\mu^*(E) \geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \cap B^C) \geq \mu^*(\cup_{j=1}^{\infty} E \cap A_j) + \mu^*(E \cap B^C) \geq \mu^*(E \cap B) + \mu^*(E \cap B^C) = \mu^*(E),$$

meaning that all inequalities are in fact equalities. So, we get

$$\sum_{n=1}^{\infty} \mu^*(E \cap A_n) = \mu^*(E \cap B).$$

If we pick $E = B$, we have

$$\sum_{n=1}^{\infty} \mu^*(A_n) = \mu^*(B) = \mu^*(\cup_{n=1}^{\infty} A_n),$$

as desired for the restriction to be a measure.

Then we show completeness. For $A \in \mathcal{A}$ with $\mu(A) = 0$ and some $B \subset A$, we know that $\mu^*(B) \leq \mu^*(A) = 0$, so $\mu^*(B) = 0$. We just need in addition, $\mu^*(B) = 0$ implies $B \in \mathcal{A}$. To show this, we consider for any subset E

$$\mu^*(E) = \mu^*(E \cap B) + \mu^*(E \cap B^C) \leq \mu^*(B) + \mu^*(E \cap B^C) = \mu^*(E \cap B^C) \leq +\mu^*(E).$$

This means B is mu^*-measurable. Done. □

This measure is good, but is it unique? Turns out, we need additional conditions to have uniqueness.

Theorem A.10. *Let $S \subset \mathcal{P}(X)$ be a semi-ring with $X \in S$ and μ_0 be a σ-finite premeasure on it. We let Σ be the set of μ^*-measurable sets. Suppose there is another outer measure $\nu^* : \mathcal{P}(X) \to [0, \infty]$ such that $\nu^* = \mu_0$ on S, then we must have $\nu^* = \mu^*$ on Σ.*

A.2 Convergence Results

Theorem A.11 (Monotone Convergence Theorem). *Let $\{f_n\}$ be a sequence of nonnegative measurable functions on E such that $f_n \leq f_{n+1}$ for every n. For every $x \in E$, we set $f(x) := \lim f_n(x)$. Then,*

$$\int f d\mu = \int \lim f_n d\mu = \lim \int f_n d\mu.$$

Proof. First, f is measurable as it is the limit (defined pointwise) of measurable functions. Next, as $f_n \leq f$, we have $\int f_n \leq \int f$ and also

$$\lim \int f_n d\mu \leq \int f d\mu.$$

So, we just need to obtain the reverse inequality to establish the desired equality.

Let us consider a simple function $h(x) := \sum_{i=1}^{k} \alpha_i 1_{A_i}$ with $h \leq f$. We consider some $a \in [0, 1)$ so $ah < f$ and we denote

$$E_n := \{x \in E \mid ah(x) \leq f_n(x)\}.$$

Since $f_n \nearrow f$, we have $E_n \nearrow E$. This gives us the following

$$\int f_n d\mu \geq \int f_n 1_{E_n} d\mu \geq \int ah(x) 1_{E_n} d\mu = a \sum_{i=1}^{k} \alpha_i \mu(A_i \cap E_n).$$

Since $E_n \nearrow E$, we would also have $A_i \cap E_n \nearrow A_i \cap E = A_i$. So, taking $\lim_n$ gives us

$$\lim \int f_n d\mu \geq \lim a \sum_{i=1}^{k} \alpha_i \mu(A_i \cap E_n) = a \sum_{i=1}^{k} \alpha_i \mu(A_i) = a \int h d\mu.$$

Since a and h are both arbitrary, we take $\lim_{a \to 1}$ and $\sup_h$, which yields

$$\lim \int f_n d\mu \geq \int f d\mu,$$

as desired. □

Theorem A.12 (Fatou Lemma)**.** *Let $\{f_n\}$ be a sequence of nonnegative measurable functions. Then,*

$$\int \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

Proof. First, we have

$$\liminf_{n\to\infty} f_n = \lim_{k\to\infty} \inf_{n\geq k} f_n.$$

Clearly, $\inf_{n\geq k} f_n$ is increasing in k, so by monotone convergence theorem, we have

$$\int \liminf_{n\to\infty} f_n d\mu = \int \lim_{k\to\infty} \inf_{n\geq k} f_n d\mu = \lim_{k\to\infty} \int \inf_{n\geq k} f_n d\mu.$$

Next, we know that $\inf_{n\geq k} f_n \leq f_p$ for any $p \geq k$. So,

$$\int \inf_{n\geq k} f_n d\mu \leq \int f_p d\mu$$

and

$$\int \inf_{n\geq k} f_n d\mu \leq \inf_{p\geq k} \int f_p d\mu.$$

Therefore, we have

$$\int \liminf_{n\to\infty} f_n d\mu = \lim_{k\to\infty} \int \inf_{n\geq k} f_n d\mu \leq \lim_{k\to\infty} \inf_{p\geq k} \int f_p d\mu = \liminf \int f_n d\mu,$$

as desired. □

Theorem A.13 (Dominated Convergence Theorem)**.** *Let $\{f_n\}$ be a sequence of integrable function, f be a measurable function with $f_n \to f$ a.e., and there exits an integrable function g such that $g \geq 0$ and $|f_n(x)| \leq g(x)$ for all x. Then, we have*

$$\int f d\mu = \int \lim f_n d\mu = \lim \int f_n d\mu.$$

Proof. Note that as $|f_n(x)| \leq g(x)$ for all n and $f_n \to f$ a.e. implies that $|f(x)| \leq g(x)$ a.e., so f is integrable as well.

Next, as $|f_n| \leq g$, we consider $g - f_n, g + f_n \geq 0$ for all n. Using Fatou, we have

$$\int g d\mu - \int f d\mu = \int \liminf_{n\to\infty}(g - f_n) d\mu \leq \liminf_{n\to\infty} \int (g - f_n) d\mu = \int g d\mu - \limsup_{n\to\infty} \int f_n d\mu$$
$$\int g d\mu + \int f d\mu = \int \liminf_{n\to\infty}(g + f_n) d\mu \leq \liminf_{n\to\infty} \int (g + f_n) d\mu = \int g d\mu + \liminf_{n\to\infty} \int f_n d\mu$$

which gives us, after rearranging the terms,

$$\limsup_{n\to\infty} \int f_n d\mu \leq \int f d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$$

and thus we have the desired equality. □

A.3 Product Measure

Earlier we have described how one could construct a measure abstractly. Here we will apply that procedure to obtain the product measure.

Consider two measure spaces $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$. A **rectangle** in $X \times Y$ is a set of the form $A \times B$ where $A \in \mathcal{M}, B \in \mathcal{N}$. It can be verified that the set of rectangles S form a semi-ring. We can defined a premeausre ρ as follows:

$$\rho(A \times B) = \mu(A)\nu(B)$$

for all $A \times B \in S$. The verification that this is indeed a premeasure involves a bit of work, but it could be done (apply MON twice). Then, we can extend it using Caratheodory and obtain a measure and a measurable set. The measure is denoted by $\pi = \mu \otimes \nu$ and the measurable set is denoted by $\mathcal{M} \otimes \mathcal{N}$.

The key result involving produce measure is the Fubini-Tonelli theorem, which gives us justification to exchange the order of integration.

We need some preliminary definitions and result first.

Definition A.14. *If $E \subset X \times Y$, then we denote $E_x := \{y \in Y \mid (x, y) \in E\}$ and $E^y := \{x \in X \mid (x, y) \in E\}$. If $f : X \times Y \to \mathbb{R}$, then we denote $f_x(y) := f(x, y)$ for any fixed $x \in X$ and $f^y(x) := f(x, y)$ for any fixed $y \in Y$.*

Proposition A.15. *Consider two measure spaces $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$.*

1. *If $E \in \mathcal{M} \otimes \mathcal{N}$, then $E_x \in \mathcal{N}$ and $E^y \in \mathcal{M}$.*
2. *If $f : X \times Y \to \mathbb{R}$ is $\mathcal{M} \otimes \mathcal{N}$ measurable, then f_x is $\mathcal{N}$ measurable and f^y is $\mathcal{M}$ measurable.*

Proof. (1) Consider the set $R := \{E \subset X \times Y \mid E_x \in \mathcal{N}, E^y \in \mathcal{M}\}$. The desired result follows if we can show $R \supset \mathcal{M} \otimes \mathcal{N}$. First, we notice that if $A \in \mathcal{M}$ and $B \in \mathcal{N}$, we have $(A \times B)_x$ is B when $x \in A$ and $\emptyset$ when $x \notin A$. Similarly $(A \times B)^y$ is A when $y \in B$ and $\emptyset$ when $y \notin B$. So, R contains all the rectangles. If we can then establish that R is an σ-algebra, then it must contain the σ-algebra generated by rectangles, which is just $\mathcal{M} \otimes \mathcal{N}$.

Notice that $(E^C)_x = (E_x)^C \in \mathcal{N}$ and $(E^C)^y = (E^y)^C \in \mathcal{M}$, which yields the closure under complement. Closure under countable union is simple too. We have $(\cup E_n)_x = \cup(E_n)_x \in \mathcal{N}$. The rest follows trivially.

(2) This is a consequence of (1). We have

$$(f_x)^{-1}(B) = (f^{-1}(B))_x$$

from definition. The rest follows from (1) trivially. □

Theorem A.16 (Fubini-Tonelli Theorem)**.** *Consider two σ-finite measure spaces $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$.*

1. *(Tonelli) Let $f \in L^+(X \times Y)$ be $\mathcal{M} \otimes \mathcal{N}$ measurable. Then the functions*

$$g(x) := \int_Y f_x(y) d\nu, \quad h(x) := \int_X f^y(x) d\mu,$$

are measurable in $L^+(X)$ and $L^+(Y)$ respectively. Also, we have

$$\int_{X \times Y} f d\pi = \int_X g d\mu = \int_Y h d\nu.$$

2. *(Fubini) Let $f \in L^1(X \times Y)$. f_x and f^y are thus in $L^1(X)$ for μ-a.e. x and $L^1(Y)$ for ν-a.e. y respectively. The functions g, h are also measurable a.e. with*

$$\int_{X \times Y} f d\pi = \int_X g d\mu = \int_Y h d\nu.$$

Index