Probability Theory

Jianing Yao
Department of MSIS-RUTCOR
Rutgers University, the State University of New Jersey
Piscataway, NJ 08854 USA

July 6, 2015

Contents

Chapter 1

Probability Space and Basics

1.1 Review of Sets Theory

In probability theory, we identify an event as a subset of outcome space. Usually, such classes of events are required to satisfy the closure properties that define *algebra* or σ*-algebra* of sets. In this section, we will review these concepts as a prologue to the theory.

To start the journey, let Ω be a non-empty set which represents the collection of all possible outcomes. We are interested in certain kinds of families of subsets of Ω.

Definition 1.1.1 Let $\mathcal{F}$ be a non-empty collection of subsets of Ω, we call $\mathcal{F}$ an *algebra* if:

(i) $A \in \mathcal{F}$ implies $A^{\mathcal{C}} \in \mathcal{F}$;
(ii) $A_1, ..., A_n \in \mathcal{F}$ implies $\cap_{i=1}^{n} A_i \in \mathcal{F}$;
(iii) $A_1, ..., A_n \in \mathcal{F}$ implies $\cup_{i=1}^{n} A_i \in \mathcal{F}$.

We call $\mathcal{F}$ a σ*-algebra* if (i) above is satisfied, in addition,

(i) $A_1, A_2, ... \in \mathcal{F}$ implies $\cup_{i=1}^{\infty} A_i$;
(ii) $A_1, A_2, ... \in \mathcal{F}$ implies $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Notice the differences lie between algebra and σ-algebra is the extension of closure property as the number of subsets grow to infinity. Therefore, any σ-algebra is an algebra and any statement that is true for algebra is automatically valid for σ-algebra. To give more insights into these two structures, we have the following result:

Proposition 1.1.1 (i) If $\mathcal{F}$ is an algebra, then Ω and $\emptyset$ belongs to $\mathcal{F}$;
(ii) if a non-empty family $\mathcal{F}$ is closed under complementation, then it is closed under countable intersections if and only if it is closed under countable unions;

(iii) If $\mathcal{F}$ is non-empty and if $A^{\mathcal{C}} \in \mathcal{F}$ and $A \cup B \in \mathcal{F}$ whenever $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $\mathcal{F}$ is an algebra;
(iv) if $\{\mathcal{F}_\alpha; \alpha \in \mathcal{I}\}$ is family of algebras (σ-algebras), where $\mathcal{I}$ is an arbitrary, non-empty index set, then $\cap_{\alpha\in\mathcal{I}}\mathcal{F}_\alpha$ is an algebra (σ-algebra).

Proof. (i) Since $\mathcal{F}$ is an algebra, for any $A \in \mathcal{F}$, $A^{\mathcal{C}} \in \mathcal{F}$. Thus, $A \cup A^{\mathcal{C}} = \Omega \in \mathcal{F}$, in the meanwhile, $A \cap A^{\mathcal{C}} = \emptyset \in \mathcal{F}$. (ii) By *De Morgan laws*:

$$\left(\bigcap_{\alpha\in\mathcal{I}} A_\alpha\right)^{\mathcal{C}} = \bigcup_{\alpha\in\mathcal{I}} A_\alpha^{\mathcal{C}} \quad \text{and} \quad \left(\bigcup_{\alpha\in\mathcal{I}} A_\alpha\right)^{\mathcal{C}} = \bigcap_{\alpha\in\mathcal{I}} A_\alpha^{\mathcal{C}}$$

Since $A_\alpha = (A_\alpha^{\mathcal{C}})^{\mathcal{C}}$, (ii) is proved. (iii) Let's use mathematical induction, the case when number of sets is two is trivial. Suppose for n sets, we have $B_n := \cup_{i=1}^n A_i \in \mathcal{F}$, then $B_{n+1} = (\cup_{i=1}^n A_i) \cup A_{n+1} \in \mathcal{F}$, thus $\mathcal{F}$ is closed under finite unions. By applying result (ii), we proved it is also closed under finite intersections. (iv) can be proved by checking definition of algebra (σ-algebra). Let's work out the case of algebra, denote $\mathcal{F}_\mathcal{I} = \cap_{\alpha\in\mathcal{I}}\mathcal{F}_\alpha$. If $A \in \mathcal{F}_\mathcal{I}$, A must belongs to $\mathcal{F}_\alpha$ for all α. But $\mathcal{F}_\alpha$ is an algebra, then $A^{\mathcal{C}} \in \mathcal{F}_\alpha$ for all α, thus $A^{\mathcal{C}} \in \mathcal{F}_\mathcal{I}$. If $A_1, ..., A_n \in \mathcal{F}_\mathcal{I}$, then $A_1, ..., A_n \in \mathcal{F}_\alpha$, since $\mathcal{F}_\alpha$ is σ-algebra, $\cup_{i=1}^n A_i \in \mathcal{F}_\alpha$ for all α, thus $\cup_{i=1}^n A_i \in \mathcal{F}_\mathcal{I}$. By (iii) in Proposition 1.1.1, $\mathcal{F}_\mathcal{I}$ is shown to be an algebra. The proof for σ-algebra proceeds exactly the same, we will leave to the reader. □

One trivial example of σ-algebra will be simply the power set of Ω. Let's check out another important example:

Example 1.1.2 Let $\mathcal{R}$ be the collection of all subintervals of the real line that are open on the left and closed on the right, i.e., $(a, b]$, it also includes (a, ∞), $(-\infty, b)$ and $(-\infty, \infty)$. Obviously, $\mathcal{R}$ is closed under finite intersections (but not finite unions). Let $\mathcal{F}$ be the collection of all finite disjoint unions of sets in $\mathcal{R}$. We claim: $\mathcal{F}$ **is an algebra**. First of all, $\mathcal{F}$ is closed under complementation. By Proposition 1.1.1, it suffices to show $A \cap B$ is in $\mathcal{F}$ whenever A and B are. Let's write

$$A = \cup_{i=1}^m R_i, \quad B = \cup_{j=1}^n S_j,$$

where $R_1, ..., R_m, S_1, ..., S_n$ are disjoint sets in $\mathcal{R}$. Then

$$A \cap B = \cup_{i=1}^n \cup_{j=1}^m R_i \cap S_j = \emptyset \in \mathcal{F}.$$

Let's generalize argument of last example:

Lemma 1.1.3 Let $\mathcal{G}$ be a collection of subsets of Ω such that:

(i) if A and B are in $\mathcal{G}$, then so is $A \cap B$;
(ii) if $A \in \mathcal{G}$, then $A^{\mathcal{C}}$ is a finite disjoint union of sets in $\mathcal{G}$;

Then the collection of all finite disjoint unions of sets in $\mathcal{G}$ is an algebra.

Proof. For any A, $B \in \mathcal{G}$, by (i) and (ii), $(A \cap B)^{\mathcal{C}} = A^{\mathcal{C}} \cup B^{\mathcal{C}}$, where $A^{\mathcal{C}}$ and $B^{\mathcal{C}}$ are both finite disjoint union of sets in $\mathcal{G}$. If we denote the collection of all finite disjoint unions of sets in $\mathcal{G}$ by $\mathcal{A}$, then the above argument tells: if $A \in \mathcal{A}$, then $A^{\mathcal{C}} \in \mathcal{A}$, by (ii) and (iii) of Proposition 1.1.1, the assertion follows. □

We have mainly shown several characterizations of algebra, however, the explicit characterization of σ-algebra is not that easy. Nevertheless, we rely on an indirect method.

Definition 1.1.2 Let $\mathcal{R}$ denote a non-empty collection of subsets of Ω. Define

$$\sigma(\mathcal{R}) := \bigcap \{ \mathcal{G} \, ; \, \mathcal{G} \text{ is a } \sigma\text{-algebra, } \mathcal{R} \subset \mathcal{G} \}.$$

Then, $\sigma(\mathcal{R})$ is called the σ-algebra generated by $\mathcal{R}$.

This definition says that the smallest σ-algebra containing the generating set is a σ-algebra, which makes sense because $\mathcal{R}$ is contained in the power set of Ω, which is a σ-algebra, and hence $\sigma(\mathcal{R})$ is not empty. If X is a topological space, the σ-algebra generated by the open sets of X is called the *Borel* σ-algebra of X and is denoted by $\mathcal{B}(X)$. We shall always view $\mathbb{R}^n$ as a topological space using the standard topology given by the distance metric. Then $\mathcal{B}(\mathbb{R}^n)$ denotes the *Borel* σ-algebra using this topology, and sets in $\mathcal{B}(\mathbb{R}^n)$ are called *Borel sets*. The *Borel sets* play an important role in *Lebesgue's theory of measure on* $\mathbb{R}^n$. The *Borel* σ-algebra of $\mathbb{R}^n$ is also generated by, for example: (i) the collection of closed subsets; (ii) the collection of open balls; (iii) the collection of all *'rectangles'* of the form $[a_1, b_1) \times \cdots \times [a_n, b_n)$; (iv) the countable collection of open balls with rational radii and center $z = (z_1, ..., z_n)$, where z_i is rational for all i.

Proof. Since each closed set is, by definition, the complement of an open set, and σ-algebra are closed under complements, the σ-algebra generated by closed sets is contained in $\mathcal{B}(\mathbb{R}^n)$. By the same principle, we can start with closed set in the σ-algebra generated by all closed subset. Then all open sets are in there as well. Then, $\mathcal{B}(\mathbb{R}^n)$ is contained in the σ-algebra generated by closed sets. This proves (i). (ii) is true because each open ball is a Borel set, and because each open set is a countable union of open balls. By the same reasoning, the generating family can be open rectangles. In terms of (iii), any right half-open rectangle can be written as an intersection of countable number of rectangles, so each right half-open rectangle is a Borel set. Conversely, any open rectangle is a countable union of right half-open rectangles. Thus the assertion follows. For (iv), obviously, the σ-algebra generated by open balls with rational center and radii is contained in the σ-algebra generated by just open balls. On the other hand, any open ball with irrational center can be realized by countable unions of open balls with irrational centers and rational radii. Furthermore, any ball with irrational center can be approximated by a sequence open balls with rational centers. Lastly, for (v) □

Another way that σ-algebras are generated in practice is by *'pullbacks'* of previously defined σ-algebras. Suppose that $\mathcal{F}$ is σ-algebra of subsets of Ω and let $f : X \mapsto \Omega$ be

a mapping from some other set X into Ω. The pre-image under f of a subset U of Ω is $f^{-1}(U) := \{\, x \in X;\, f(x) \in U \,\}$. Then

$$f^{-1}(\mathcal{F}) := \{\, f^{-1}(U);\, U \in \mathcal{F} \,\},$$

defines a σ-algebra of subsets of X. This fact is a consequence of the identities:

$$f^{-1}\left(\bigcap_{\alpha\in\mathcal{I}} U_\alpha\right) = \bigcap_{\alpha\in\mathcal{I}} f^{-1}(U_\alpha) \tag{1.1}$$

and

$$f^{-1}\left(\bigcup_{\alpha\in\mathcal{I}} U_\alpha\right) = \bigcup_{\alpha\in\mathcal{I}} f^{-1}(U_\alpha), \tag{1.2}$$

Proof. For any $A_i \in f^{-1}(\mathcal{F})$, it is represented by $f^{-1}(U_i)$, for some $U_i \in \mathcal{F}$, where $i \in \mathcal{I}$. For $A^\infty = \bigcap_{\alpha\in\mathcal{I}} A_i$, by (1.1),

$$A^\infty = \bigcap_{\alpha\in\mathcal{I}} f^{-1}(U_\alpha) = f^{-1}\left(\bigcap_{\alpha\in\mathcal{I}} A_\alpha\right) \in f^{-1}(\mathcal{F})$$

By the same arguments, the closure under countable unions can be verified. Also, for any $A = f^{-1}(U)$, $U \in \mathcal{F}$,

$$A^{\mathcal{C}} = \left(f^{-1}(U)\right)^{\mathcal{C}} = f^{-1}(U^{\mathcal{C}}) \in f^{-1}(\mathcal{F})$$

Thus, $f^{-1}(\mathcal{F})$ is a σ-algebra. □

Projection maps on product space provide a simple example.

Example 1.1.4 Let $\pi_1(x_1, ..., x_n) = x_1$ denotes the projection of $\mathbb{R}^n$ on its first coordinate. Then, $\pi_1^{-1}(\mathcal{B}(\mathbb{R}))$ is a σ-algebra and it may be explicitly characterized as the collection of all subsets of $\mathbb{R}^n$ of the form $U \times \mathbb{R}^{n-1}$, where U is a *Borel* subset of $\mathbb{R}$. With this result, we can prove that $\bigcup_{k=1}^n \pi_k^{-1}(\mathcal{B}(\mathbb{R}))$ is the generator of $\mathcal{B}(\mathbb{R}^n)$. Since $\pi_k^{-1}(\mathcal{B}(\mathbb{R})) = U \times \mathbb{R}^{n-1}$, where $U \in \mathcal{F}$, is contained in $\mathcal{B}(\mathbb{R}^n)$ for all k. The finite union of $\pi_k^{-1}(\mathcal{B}(\mathbb{R}))$ is contained in the sigma algebra generated, thus $\sigma\left(\bigcup_{k=1}^n \pi_k^{-1}(\mathcal{B}(\mathbb{R}))\right) \subseteq \mathcal{B}(\mathbb{R}^n)$. On the other hand, $\mathcal{B}(\mathbb{R}^n)$ is generated by open sets in $\mathbb{R}^n$, but all such open sets in contained in $\bigcup_{k=1}^n \pi_k^{-1}(\mathcal{B}(\mathbb{R}))$, thus, we have the inverse inclusion. The assertion follows.

1.2 Probability Space

When modelling a random experiment, such as, rolling a die or observing a particle undergoing Brownian motion. The experiment will have a set of possible outcomes, which is

usually denoted by Ω. It is called the *outcome space*. For example, for one roll of a die, $\Omega = \{\, 1,\, 2,\, 3,\, 4,\, 5,\, 6 \,\}$, for Brownian motion, Ω is a set of continuous functions of time with values in $\mathbb{R}^3$. An *event*, in the colloquial sense, is not the same as an outcome. For example, we speak of the event of getting an even number in the roll of a die, or of the event that a Brownian motion moves more than a centimeter from its starting point in on second. In each case, the event is defined by a subset of Ω: an even roll occurs if the outcome falls in the subset $\{\, 2,\, 4,\, 6 \,\}$ of $\{\, 1,\, 2,\, 3,\, 4,\, 5,\, 6 \,\}$; a Brownian motion path moves more than one centimeter over a second if it belongs to the continuous functions, f, for which $|f(1) - f(0)| > 1$. Given a subset A of Ω, we say A occurs in a random trial, if the outcome falls in A.

The object of modelling is to assign probabilities, $\mathbb{P}(A)$ to subsets A of Ω. Here, $\mathbb{P}(A)$ is a number between 0 and 1 that represents the probability that A occurs, where a probability of 1 means the event is certain to occur, and a probability of 0 means it will not occur. We may not want to assign probabilities to all subsets of Ω and, as we shall see, in fact, it may not even be possible to do so in meaningful way. We shall denote the class of subsets on which $\mathbb{P}$ is defined by $\mathcal{F}$. By the term event, we shall mean especially a subset of Ω that belongs to $\mathcal{F}$, not any subset of Ω.

The only general constraint imposed on $\mathbb{P}$ will be an additivity condition: the probability of a union of disjoint sets should be the sum of the individual probabilities. This is entirely consistent with intuition, whether from the *frequentist* or *Bayesian* perspective. Formal definitions of probability space follow, there are two kinds:

Definition 1.2.1 The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a finitely additive probability space if Ω is a non-empty set, $\mathcal{F}$ is an algebra of subsets of Ω, and $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ is a *finitely additive measure* on $\mathcal{F}$ satisfying $\mathbb{P}(\Omega) = 1$. Finite additivity means:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i), \;\; \textit{whenever } A_1, ..., A_n \textit{ in } \mathcal{F} \textit{ are disjoint.}$$

The assumptions of this definition are reasonable, minimal conditions to put on any probability model. However, finitely additivity is not powerful enough to treat the limit questions that are ubiquitous in probability theory and require countable set operations. For this, we make a further definition:

Definition 1.2.2 The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space if Ω is a non-empty set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $\mathbb{P}$ is a non-negative function on $\mathcal{F}$ satisfying $\mathbb{P}(\Omega) = 1$ and

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i), \;\; \textit{for any disjoint family } \{A_1, A_2, ...\} \subset \mathcal{F}.$$

Remark 1.2.1 In the language of measure theory, $\mathbb{P}$ is a countably additive, positive measure on $(\Omega, \mathcal{F})$ of total mass one.

This set of conditions placed on $(\Omega, \mathcal{F}, \mathbb{P})$ are called *axioms of probability*. Notice, the term *'probability space'*, without further qualification, always means a countably additive space. Of course, a probability space is automatically a finite additive probability space, and, when $\mathcal{F}$ is finite, countable additivity and finite additivity are equivalent. But when we take trouble to say a probability space is finitely additive, we usually mean either that $\mathcal{F}$ is not a σ-algebra, or, if it is, that $\mathbb{P}$ is not countably additive.

Most text do not specially define finitely additive probability spaces. We do so here for two reasons: firstly, when Ω is uncountable, probability space are usually constructed by first defining a finitely additive probability space and then extending it; secondly, finitely additive probability spaces that are not countably additive do sometimes arises as models.

Remark 1.2.2 Finitely additivity is a simple property supported directly by intuition. Countable additivity is a stronger demand that is not so immediately intuitive. Are we justified in assuming it as the basis of probability theory for more than just reasons of convenience? As we indicated, countably additive probability spaces are often constructed by extending finitely additive probability spaces. This is done by using *Carathéodory's extension theorem* and requires only checking that the finitely additive probability space has a certain continuity property. When it odes, the extension is unique.

Example 1.2.3 *(Probability space for the roll of a fair die)* Let $\Omega := \{ 1, 2, 3, 4, 5, 6 \}$, let $\mathcal{F}$ be the collection of all subsets of Ω, and let $\mathbb{P}(A) := |A|/6$, for every $A \subset \Omega$, where $|A|$ denotes the cardinality of A. This is a model for a fair die, because for each $i \in \{ 1, ..., 6 \}$, the event of rolling i is $\{i\}$ and $\mathbb{P}(\{i\}) = 1/6$. It is easy to show that $\mathbb{P}$ is finitely additive, and hence, since $\mathcal{F}$ is finite, countably additive.

Example 1.2.4 *(Finite outcome space with equally likely outcomes)* This example generalizes the above example. If A is a finite set, let $|A|$ denote its cardinality again. Let the outcome space Ω be finite. Let $\mathcal{F}$ be the power set of Ω, that is the collection of all subsets of Ω, and define

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

It is easy to check $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. For any $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) = 1/|\Omega|$, and thus each possible outcome is equally likely. For this reason, $\mathbb{P}(A)$ is called the *uniform probability distribution on* Ω.

Example 1.2.5 *(Discrete probability spaces)* This example subsumes all the previous ones as special cases. Let the outcome space Ω be a countable, and let $\mathcal{F}$ be the power set of Ω. Suppose that for each $\omega \in \Omega$, we have a number $p_\omega \geq 0$ and that

$$\sum_{\omega \in \Omega} p_\omega = 1.$$

Define $\mathbb{P}(A) = \sum_{\omega \in A} p_\omega$, for any $A \in \Omega$. It is an easy exercise to show that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. In this model, $p_\omega = \mathbb{P}(\{\omega\})$ is the probability that the outcome is ω. Elementary combinatorial probability is a study of probability space of this kind.

Example 1.2.6 *(Uniform distribution on* $[0,1]$*, finitely additive version)* Imagine an experiment whose outcome is a number in $[0,1]$, that, heuristically speaking, is equally likely to fall anywhere. More precisely, the probability that the random point fall into a subset A should be the same for any translation (modulo 1) of A. Also, it is natural to want any interval to be an event. To define a finitely additive probability space with these properties, let $\mathcal{G}$ be the collection of all finite disjoint unions of subintervals of $[0,1]$, and, for any $A \in \mathcal{G}$, let $l(A)$ be the total length of A. Then, $\mathcal{G}$ is an algebra. The length measure l is finitely additive, and it is invariant under translation. Hence, $([0,1], \mathcal{G}, l)$ is a finitely additive probability space modelling a point uniformly selected from $[0,1]$.

Example 1.2.7 *(Uniform distribution of* $[0,1]$*, countably additive version)* This example assumes knowledge of *Lebesgue measure*. However, we shall explain the construction fully later on. Let $\Omega = [0,1]$ and let $\mathcal{F} = \mathcal{B}([0,1])$, the σ-algebra of the Borel subsets of $[0,1]$. Let λ denote Lebesgue measure. since λ is a countably additive and $\lambda([0,1]) = 1$, the triple $([0,1], \mathcal{B}([0,1]), \lambda)$ is a probability space. The measure λ is also invariant under translation of $[0,1]$ modulo 1. Also $\sigma(\mathcal{G}) = \mathcal{B}([0,1])$, where $\mathcal{G}$ is the algebra defined in above example, and $\lambda(A) = l(A)$ for every $A \in \mathcal{G}$. We say that countably additive model $([0,1], \mathcal{B}([0,1]), \lambda)$ extends the finitely additive model of above example. As we shall see, countably additive probability spaces are often constructed as extension of finitely additive spaces.

If we assume the axiom of choice, which we do in this text, there is no countably additive, translation invariant, probability measure on all subsets of $[0,1]$. This is a consequence of the well known construction of sets which are not Lebesgue-measurable. Most any graduate real analysis text will have a proof of this result. It illustrates why it is not always possible to insist that $\mathcal{F}$ be the power set of Ω.

The last example has an annoying feature it shares with most probability models on uncountable outcome spaces. Although the element ω of Ω are the possible outcomes, $\mathbb{P}(\{\omega\}) = 0$ for every ω. So every time the experiment runs, an outcome of probability zero occurs! This paradox does not mean the uniform distribution is useless. In real life, no measurement is arbitrarily accurate. If measurements are accurate only to $1/N$, the experiment is really returning one of the numbers in the finite set $\{\, 0,\ 1/N,\ 2/N,\ ...,\ 1 \,\}$. Suppose each of these points is equally probable. When N is large, the uniform distribution on $[0,1]$ will be a good approximation to this discrete model, and one that is easier to work with.

1.3 Elementary properties of probability measures

In this section, we develop some immediate and elementary consequences of axioms of probability.

Theorem 1.3.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a *finitely additive* probability space. Then

(i) $\mathbb{P}(\emptyset) = 0$;
(ii) if $A \in \mathcal{F}$, $\mathbb{P}(A^{\mathcal{C}}) = 1 - \mathbb{P}(A)$;
(iii) if $A \subset B$ and $A, B \in \mathcal{F}$, then $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$, and $\mathbb{P}(A) \leq \mathbb{P}(B)$;
(iv) if $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
(v) If $A_1, ..., A_n \in \mathcal{F}$,

$$\mathbb{P}\left(A_1 \cup \cdots \cup A_n\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 \leq \cdots i_r \leq n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_r}\right)$$

(vi) $\mathbb{P}$ is *sub-additive*: $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$.

The proof will be skipped here, just to mention that properties (i)-(iv) are elementary and the generalized inclusion-exclusion identity can be proved by induction. Property (iv) is the special case of the inclusion-exclusion principle for $n = 2$. Property (vi) follows from (iv) and induction.

The next result relates countable additivity to *"continuity"* properties of a probability measure.

Theorem 1.3.2 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

1. $\mathbb{P}$ is continuous from below; that is, if A_n is an increasing sequence of events,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

2. $\mathbb{P}$ is continuous from above; that is, if A_n is decreasing sequence of events,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

Proof. Let $A_1 \subset A_2 \subset \cdots$ be an increasing sequence of events. Define $B_1 = A_1$ and for $n \geq 2$, define $B_n = A_n - A_{n-1}$. Then B_1, B_2,... are disjoint, $\bigcup_{i=1}^{n} B_i = A_n$ for every $n \geq 1$ and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Therefore, using countable additivity,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

This proves (i).

To prove (ii), use the fact that $(\bigcap_{n=1}^{\infty} A_n)^{\mathcal{C}} = \bigcup_{n=1}^{\infty} A_n^{\mathcal{C}}$. If the sequence $A_1, A_2, ...$ is decreasing, then $A_1, A_2, ...$ is increasing. Thus from (i),

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^{\mathcal{C}}\right) = \lim_{n \to \infty} 1 - \mathbb{P}(A_n^{\mathcal{C}}) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

□

Let $A_n \downarrow A$ means that $\{A_n\}$ is decreasing sequence of sets and $\bigcap_{n=1}^{\infty} A_n = A$. Let P be a positive function on a family, $\mathcal{C}$, of subsets of Ω. We say that P on $\mathcal{C}$ is continuous from above at $\emptyset$, if $\lim_{n\to\infty} \mathbb{P}(C_n) = 0$ whenever $\{C_n\}$ is a sequence of sets in $\mathcal{C}$ such that $C_n \downarrow \emptyset$. Theorem 1.3.2 shows that countable additivity of $\mathbb{P}$ implies continuity form above of $\mathbb{P}$ and thus from above at $\emptyset$. The next theorem is kind of a converse statment.

Theorem 1.3.3 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a *finite additive* probability space and assume in addition that $\mathcal{F}$ is a σ-algebra. If $\mathbb{P}$ on $\mathcal{F}$ is continuous from above at $\emptyset$, then $\mathbb{P}$ is countably additive on $\mathcal{F}$.

Proof. Let $B_1, B_2, ...$ be a disjoint sequence of events. Let $A_n = \bigcup_{i=n}^{\infty} B_n$. Then $A_n \downarrow \emptyset$ and so $\mathbb{P}(A_n) \downarrow 0$ as $n \to \infty$. By finite additivity of $\mathbb{P}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) + \mathbb{P}(A_{n+1}) = \sum_{i=1}^{n} \mathbb{P}(B_i) + \mathbb{P}(A_{n+1}),$$

for every n. By letting $n \to \infty$, $\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i)$. □

1.4 The Extension Problem

If $(\Omega, \mathcal{C}, P)$ is a finitely additive probability space and if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space such that

$$\mathcal{C} \subset \mathcal{F} \text{ and } \mathbb{P}(A) = P(A) \text{ for all } A \in \mathcal{C}$$

then we say that $(\Omega, \mathcal{F}, \mathbb{P})$ is a *countably additive extension* of P from $(\Omega, \mathcal{C})$ to $(\Omega, \mathcal{F})$. When this is the case, $\sigma(\mathcal{C}) \subset \mathcal{F}$, so $(\Omega, \sigma(\mathcal{C}), \mathbb{P})$ is also an extension of $(\Omega, \mathcal{C}, P)$.

The question arises: under what conditions does a finitely additive probability space $(\Omega, \mathcal{C}, P)$ admit a countably additive extension? The answer is fairly simple. Observe from Theorem 1.3.3 that if P admits such an extension, then P on $\mathcal{C}$ must be continuous from above at $\emptyset$. In fact, this is also a sufficient condition.

Theorem 1.4.1 *(Carathéodory's extension theorem)* Let $(\Omega, \mathcal{C}, P)$ be a finitely-additive probability space. Then there is a countably additive measure $\mathbb{P}$ on $(\Omega, \sigma(\mathcal{C}))$ that extends P on $(\Omega, \mathcal{C})$ if and only if P on $\mathcal{C}$ is continuous from above at $\emptyset$. When it exists, this extension is unique.

This theorem is actually the special case for bounded measures of a more general result. The proof of uniqueness of the extension is deferred to later sections by an application of the *Monotone Class Theorem*. Theorem 1.4.1 is much deeper than Theorem 1.3.3. An outline of the construction of $\mathbb{P}$ can be found in the appendix. For a more general statement and a full proof, see the references at the end of this chapter.

Carathéodory's extension theorem is the main tool for constructing probability space on uncountable outcome spaces. When modelling, it is usually easy to construct a finitely additive probability space very explicitly. To derive probability space model, one then tries to show continuity from above at $\emptyset$ for the finitely additive space. This is usually a more difficult step. Often Ω carries a natural metric, and the proof will use approximation by sets that are compact in the metric topology. We will show how this works in examples.

1.5 Random Variables

Before going to the concrete examples, we may give a brief introduction of random variables for the sake of illustration of our examples. We will discuss more in details in later chapters.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $X : \Omega \mapsto \mathbb{R}$ be a real-function on Ω. In terms of modelling, we can think of $X(\omega)$ as some numerical attribute of the outcome ω. For example, in the study of Brownian motion, ω represents a continuous path. We might be interested not in the whole path, but in its value, $X(\omega) = \omega(t)$, at some time t, or its maximum up to time t, $Y(\omega) = \max\{\, \omega(s); 0 \leq s \leq t \,\}$.

Given such an X, the simplest thing we might ask for is the probability that $X(\omega) > a$ for some a. To answer this, the set $X^{-1}((a, \infty)) = \{\, \omega; X(\omega) \in (a, \infty) \,\}$ needs to be an event, that is, it must belong to $\mathcal{F}$, for any a. Suppose this is true. The it follows that

$$X^{-1}(U) \in \mathcal{F} \text{ for all Borel subsets } U \text{ of } \mathbb{R}. \tag{1.3}$$

It says the image of X can be all Borel subsets of $\mathbb{R}$ not just sets of the format (a, ∞). To see why, let $\mathcal{C} = \{\, U \subset \mathbb{R}; X^{-1}(U) \in \mathcal{F} \,\}$. It is easy to verify that this is σ-algebra. If $\mathcal{C}$ contains all set of the form (a, ∞), then, since such sets generate the Borel σ-algebra, $\mathcal{C}$ must contain the Borel sets, and the representation (1.3) is valid.

Now recall that in measure theory, we say that a function $X : (\Omega, \mathcal{F}) \mapsto \mathbb{R}$ is *Borel*-measurable if it satisfies (1.3). In probability space, we call a Borel measurable function a *random variable*. For the generic countable probability space, in which every subset of Ω is in $\mathcal{F}$, any function on Ω is measurable. This is not true for uncountable outcome space, assuming the axiom of choice, because of the existence of non-measurable sets.

Definition 1.5.1 Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the σ-algebra:

$$\sigma(X) := X^{-1}(\mathcal{B}(\mathbb{R}))$$

is called the *σ-algebra generated by X*.

Remark 1.5.1 We may think of it as the class of all events concerning the outcome of X.

1.6 Examples of Probability Space

1.6.1 Finite tosses of a fair coin

Let a coin be flipped N times. If we record each head by the number 1 each tail by the number 0, an outcome of this experiment can be represented as a sequence $\omega = (\omega_1, ..., \omega_N)$ of 1's and 0's with length N. The outcome space is then the set $\Omega_n = \{0, 1\}^N$ of all such sequences.

We want to model the case in which the coin is fair and no flips affect the outcomes of other flips (independence). Intuitively, this means that every possible sequence of flips should be equally likely. It makes sense to let the σ-algebra for our model be the class, $\mathcal{F}_N$ of all subsets of Ω_N. Then the appropriate probability measure is:

$$\mathbb{P}_N(B) := \frac{|B|}{2^N}, \text{ for } B \subset \{0, 1\}^N.$$

In particular, the probability of any specific sequence of heads and tails is $1/2^N$. Here, B is certain event. For example, an easy problem from elementary probability is to find the probability that there are exactly m heads in N tosses. This is represented by the set of all sequences in $\{0, 1\}^N$ with exactly m 1's. The cardinality of this set is the number of ways to choose m from N. Hence,

$$\mathbb{P}_N(\ m \text{ heads in } N \text{ tosses }\ ) = C_N^m \frac{1}{2^N}.$$

For each $1 \leq i \leq N$, the projection $X_i^N((\omega_1, ..., \omega_N)) = \omega_i$ of Ω_N on the i-th coordinate is a random variable; X_i^N simply records the result of the i0th toss. It is an easy, but worthwhile first exercise in thinking about σ-algebras generated by random variables, to confirm that $\sigma(X_i^N)$ consists of four sets: the empty set, Ω_N, the event $\{\, \omega; X_i^N(\omega) = \omega_i = 1 \,\}$ that toss i results in heads, and the event $\{\, \omega; X_i^N(\omega) = \omega_i = 0 \,\}$ that toss i results in tails.

If this model is to make sense, it should be consistent over different values of N. For example, consider the event that there are two heads in the first three tosses. We can consider this as an event in any of the probability spaces $(\{\, 0,\, 1 \,\}^N, \mathcal{F}_N, \mathbb{P}_N)$, where $N \geq 3$. Consistency means that the probability of the event should be the same no matter what N we use to compute it. This is clearly true because, whatever the choice of N, all outcomes concerning the first three flips alone are equally likely, just as for $N = 3$.

For a more formal and genera proof, consider a subset B of $\{\, 0,\, 1 \,\}^N$. Then,

$$\{\, (\omega_1, ..., \omega_M);\, (\omega_1, ..., \omega_N) \in B \,\} = B \times \{\, 0,\, 1 \,\}^{M-N}$$

expresses the same event that B represents as a subset of $\{\, 0,\, 1 \,\}^M$. Consistency requires:

$$\mathbb{P}_N(B) = \mathbb{P}_M(B \times \{\, 0,\, 1 \,\}^{M-N})$$

where B is a subset of $\{0, 1\}^N$ and $M > N$. But this is easy to show because $|B \times \{0, 1\}^{M-N}| = |B|2^{M-N}$, and, hence the right-hand side is:

$$\frac{|B \times \{0, 1\}^{M-N}|}{2^M} = \frac{|B|2^{M-N}}{2^M} = \frac{|B|}{2^N} = \mathbb{P}_N(B).$$

1.6.2 Infinite tosses of a fair coin

In this part, we construct a probability space for a fair coin tossed an infinite number of times in succession. Alternatively, we could imagine a countably infinite number of coins all tossed at once. Of course, such experiments could not really be performed. But we wish to have a framework for analyzing the behaviour of a sequence of coin tosses, in the limit as the number of tosses goes to infinity. For instance, will the empirical frequency of heads in N tosses converge with probability to $1/2$, as we should hope if our model is to be consistent with the frequency interpretation of probability? Even to formulate this question requires a probability space for an infinite sequence of tosses.

Let's again record each head by 1 and each tail by 0. The outcome space will be the product space $\Omega := \{0, 1\}^\infty$, that is, the set of all sequences $\omega = (\omega_1, \omega_2, ...)$ whose entries are each either 0 or 1. Here, the component ω_i represents the outcome of the i-th toss. Again, let $X_i(\omega) = \omega_i$ represent the outcome of toss i.

To construct a probability space we shall impose only two requirements. First, every subset of Ω defined by the outcomes of a finite number of tosses should be an event. Equivalently, each X_i must be a random variable. Second, the probability assigned to any event depending only on a finite number of tosses should be consistent with the finite toss model of the previous section. First we will show these two requirements will completely determine a finitely additive probability space. Then, we show this finitely additive measure is continuous from above at $\emptyset$. Carathéodory's extension theorem then gives us a countably additive probability space.

The first step is a formal definition of the class of events defined by a finite number of tosses. A subset A of $\{0, 1\}^\infty$ is called a *cylinder set* if there is a positive integer N and a subset B of $\{0, 1\}^N$ such that

$$A = \{\omega \in \{0, 1\}^\infty; (\omega_1, ..., \omega_N) \in B\}$$

This is called a *cylinder set* because $A = \pi_N^{-1}(B)$, where π_N is the projection of $\{0, 1\}^\infty$ onto $\{0, 1\}^N : \pi_N((\omega_1, \omega_2, ...)) = (\omega_1, ..., \omega_N)$. Another way to write A is $A = B \times \{0, 1\}^\infty$. It represents the event B as a subset of infinite coin toss space.

Let $\mathcal{C}$ denote the collection of all cylinder sets. We claim that $\mathcal{C}$ is an algebra. Clearly if A is an event depending only on the outcomes of the first N tosses, so is $A^\mathcal{C}$ If $A_1, ..., A_n$ are cylinder sets and A_i depends only on the first N_i tosses for each i, then $\bigcup_{i=1}^N A_i$ depends only on the first $N = \max_{1 \le i \le n} N_i$ tosses, so it is again a cylinder set.

For somewhat more formal perspective that will be useful, let $\mathcal{C}_N$ be the collection of all cylinder sets depending at most on what happens in the first N tosses:

$$\mathcal{C}_N := \{\, \pi_N^{-1}(B); B \subset \{\, 0,\, 1 \,\}^N \,\}$$

Then,

$$\mathcal{C} = \bigcup_{N=1}^{\infty} \mathcal{C}_N$$

Each $\mathcal{C}_N$ is a finite algebra, and $\mathcal{C}_N \subset \mathcal{C}_{N+1}$ for all N. Since the union of an increasing sequence of algebra is an algebra. However, $\mathcal{C}$ is not a σ-algebra. Any singleton set $\{\omega\}$ is a decreasing intersection of sets in $\mathcal{C}$, but it does not belong to $\mathcal{C}_N$ for any finite N.

The second step is to assign a probability $P(A)$ to each cylinder set. If A is a cylinder set, there is an N and a B in $\{\, 0,\, 1 \,\}^N$ such that $A = \pi_N^{-1}(B)$; A is just the event B embedded into the space of an infinite sequence of tosses. Consistency with the finite toss model then demands that:

$$P(A) = P(\pi_N^{-1}(B)) = \mathbb{P}_N(B) = \frac{|B|}{2^N}.$$

Since N and B such that $A = \pi_N^{-1}(B)$ are not unique, we must make sure that this definition is consistent. But we proved this above, because we showed that the finite toss models are consistent over different values of N.

Theorem 1.6.1 The triple $(\{\, 0,\, 1 \,\}^\infty, \mathcal{C}, P)$ is a finitely additive probability space.

Proof. We only need to show finite additivity, since $\mathcal{C}$ is already shown as an algebra. If $A, B \in \mathcal{C}$, then, since the sequence $\{\mathcal{C}_N\}$ is increasing, there is an N such that A and B are both in $\mathcal{C}_N$. Write $A = \pi_N^{-1}(A_N)$ and $B = \pi_N^{-1}(B_N)$, where $A_N, B_N \in \mathcal{C}_N$. If A and B are disjoint, then so are A_N and B_N and $A_N \cup B_N = \pi_N^{-1}(A_N \cup B_N)$. Thus, since $\mathbb{P}_N$, as defined for the finite coin toss space, is finitely additive,

$$P(A \cup B) = \mathbb{P}_N(A_N \cup B_N) = \mathbb{P}(A_N) + \mathbb{P}(B_N) = P(A) + P(B).$$

□

Theorem 1.6.2 P is continuous from above at $\emptyset$ on $\mathcal{C}$. Therefore, by *Carathédory's extension theorem*, there is a unique probability measure $\mathbb{P}$ on $(\{\, 0,\, 1 \,\}^\infty, \sigma(\mathcal{C}))$ that extends P on $(\{\, 0,\, 1 \,\}^\infty, \mathcal{C})$.

This theorem gives us our countable additive probability space for an infinite coin tosses. It is a direct consequence of the following result:

Lemma 1.6.3 Let $\{A_n\}$ be a *decreasing* sequence of *cylinder sets* of $\{\, 0,\, 1 \,\}^\infty$ for which $A_n \downarrow \emptyset$. Then there is an N such that $A_n = \emptyset$, for all $n \geq N$.

Then the proof of Theorem 1.6.2 follows easily. If $\{A_n\}$ is decreasing sequence of cylinder sets and $A_n \downarrow \emptyset$, then $A_n = \emptyset$ when $n \geq N$ for some N, and hence $P(A_n) = 0$ for all $n \geq N$.

Remark 1.6.4 Lemma 1.6.3 is actually a special case of an important theorem of point-set topology called *Tychonoff's theorem*. This says that the product of any family of compact topological spaces is compact with respect to the product topology. To obtain the lemmas from *Tychonoff's theorem*, put the discrete topology on $\{0, 1\}$ (notice all subsets of it are open and hence all subsets are closed.) By *Tychonoff's theorem*, the product $\{0, 1\}^\infty$ itself is compact. Since every cylinder set is closed, it is hence compact in the product topology (any closed subset of compact set is compact). In general, if every set in a decreasing sequence of compact sets is non-empty, then the intersection of all the sets is non-empty (result from real analysis). Lemma 1.6.3 follows immediately.

For readers unfamiliar with *Tychonoff's theorem*, we give a direct proof.

Proof. We shall prove the contrapositive: if $\{A_n\}$ is a decreasing sequence of cylinder sets and A_n is non-empty for all n, then $\cap_n A_n$ is non-empty. It will be convenient to define a notion of limit in $\{0, 1\}^\infty$. If $\{\omega^{(n)}\}$ is a sequence of points in $\{0, 1\}^\infty$, $\lim_{n\to\infty} \omega^{(n)} = \omega$ means $\lim_{n\to\infty} \omega_i^{(n)} = \omega_i$ for all i. Since $\omega_i^{(n)}$ can take only the values 0 and 1, this is the same as demanding that for each i, there is an integer M, possibly depending on i, such that $\omega_i^n = \omega_i$ for all $n \geq M$.

The following fact will be used. Let A by a cylinder set and suppose $\omega^{(n)} \in A$ for all n and $\lim_{n\to\infty} \omega^{(n)} = \omega$. Then $\omega \in A$ also. Indeed, let N be large enough so that $A \in \mathcal{C}_N$. There will be a k large enough so that $\omega_i^{(n)} = \omega_i$ for all $1 \leq i \leq N$ when $n \geq k$. Thus, $\pi_N(\omega) = \pi_N(\omega^{(n)})$ for all $n \geq k$, and hence $\omega \in A$.

Now assume $\{A_n\}$ is a decreasing sequence of cylinder sets, each of which is non-empty, and let $\omega^{(n)}$ be in A_n for each n. We will use a diagonalization procedure to construct a subsequence $\{\omega^{(n_r)}\}$ which converges. This will complete the proof. Indeed, suppose we have a convergent subsequence $\{\omega^{(n_r)}\}$, and let $\omega = \lim_{r\to\infty} \omega^{(n_r)}$. Fix any n, since $\omega^{(m)} \in A_n$ for all $m \geq n$, it follows from the previous paragraph that $\omega \in A_n$. Since this is true for all n, $\omega \in \bigcap A_n$, proving the intersection is non-empty.

It remains to prove the existence of a converging subsequence of $\{\omega^{(n)}\}$. Each component $\omega_i^{(n)}$ of $\omega^{(n)}$ is either 0 or 1. Therefore, there is a subsequence $\{n_{1k}; k \geq 1\}$, such that $\lim_{k\to\infty} \omega_1^{(n_{1k})}$ exists; call this limit ω_1. Likewise, there is a subsequence $\{n_{2k}; k \geq 1\}$ of $\{n_{1k}\}$ such that $\omega_2 := \lim_{k\to\infty} \omega_2^{(n_{2k})}$ exists. Continuing recursively, for each $j \geq 2$ there is a sequence $\{n_{jk}; k \geq 1\}$ that is a subsequence of $\{n_{j-1,k}; k \geq 1\}$ and for which $\omega_j = \lim_{k\to\infty} \omega_j^{(n_{jk})}$ exists. Then $\omega = (\omega_1, \omega_2, ...)$ defines a point in $\{0, 1\}^\infty$. Consider $\{n_{kk}; k \geq 1\}$. This is a subsequence of $\{n_{jk}; k \geq 1\}$ for all j. It follows that $\lim_{k\to\infty} \omega_j^{n_{kk}} = \omega_j$ for all j, and thus $\lim_{k\to\infty} \omega^{n_{kk}} = \omega$. □

The probability space we just constructed allows us to pose and discuss the law of large numbers for tosses of a fair coin. We present this here as an introduction to a major theme of probability theory.

Given $\omega = (\omega_1, \omega_2, ...) \in \{ 0, 1 \}^\infty$, let $X_i(\omega) = \omega_i$ be the outcome of toss i, as before. Then

$$\frac{1}{n} \sum_{i=1}^{n} X_i(\omega)$$

is the *empirical frequency* of heads in the first n tosses of ω. The frequency interpretation of probability assumes that this frequency tends to the probability of heads as $n \to \infty$. In probability, this is called a *law of large numbers*.

Is the law of large numbers true in $\{ 0, 1 \}^\infty$ true in $(\{ 0, 1 \}^\infty, \sigma(\mathcal{C}), \mathbb{P})$, the space we just constructed? If not, the most natural and simplest coin tossing model is not consistent with the frequency interpretation of probability. Fortunately, we have the following result:

Proposition 1.6.5 The set $\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{2} \}$ is an event in $\mathcal{C}$ and

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{2} \right) = 1$$

This proposition is actually a special case of a much more general law of large numbers that will be proved in later chapters. If we were being more careful, we would write the limit event above as:

$$\{ \omega \,;\, \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \frac{1}{2} \}$$

In probability theory, one typically does not display the dependence of a random variable on ω in the underlying probability space.

1.6.3 A general result about coin tossing measures

The reader might have noticed that no special property of P was used in the proof in Theorem 1.6.2 that P has a countably additive extension. The proof works for any finitely additive measure on $(\{ 0, 1 \}^\infty, \mathcal{C})$. Therefore, we obtain the following theorem:

Theorem 1.6.6 Any finitely additive probability measure on $(\{ 0, 1 \}^\infty, \mathcal{C})$ admits a *unique extension* to a countably additive probability measure on $(\{ 0, 1 \}^\infty, \sigma(\mathcal{C}))$.

1.6.4 Uniform measure on $[0,1]$

We now show how to construct the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$ of Example 1.2.7. We claimed that this probability space is an extension of the finitely additive probability space defined in Example 1.2.6. However, rather than apply Carathéodory's theorem to Example 1.2.6, we shall use a simpler finitely additive space. The argument is standard from real analysis.

Let $\mathcal{G}$ be the collection of all finite, disjoint union of subintervals of the form $(a,b]$, or $[0,b]$, where $0 \le a \le b \le 1$. The reader should check that $\mathcal{G}$ is an algebra and $\sigma(\mathcal{G}) = \mathcal{B}([0,1])$. Let l be the length measure. More precisely, if $A \in \mathcal{G}$, let

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A; \\ 0 & \text{otherwise,} \end{cases}$$

be the indicator function of A, and define $l(A)$ as the *Riemann integral*:

$$l(A) = \int_0^1 \mathbf{1}_A(x)dx.$$

This makes sense because $\mathbf{1}_A$ is piecewise constant and, since it is in $\mathcal{G}$, has only a finite number of discontinuities. Then it is easy to see that l is finitely additive on $\mathcal{G}$.

Lemma 1.6.7 *l on $\mathcal{G}$ is continuous from above at $\emptyset$.*

Using this Lemma and Carathéodory's extension theorem, we now define λ to be the unique extension of l to $([0,1], \mathcal{B}([0,1]))$. This completes the construction of $([0,1], \mathcal{B}([0,1]), \lambda)$.

Proof. (Lemma 1.6.7) The crucial observation is that if $A \in \mathcal{G}$ and $\epsilon > 0$, there is a $B \in \mathcal{G}$, such that the closure $\mathrm{cl}(B)$ of B is contained in A and $l(A) < \epsilon + l(B)$. To prove this, it is clear enough to show it is true for any A of the form $(a,b]$. But if $a < a' < b$,

$$l\left((a,b] - (a',b]\right) = l\left((a,a']\right) = a' - a.$$

Thus, if $a < a' < a + \epsilon$, $l((a,b]) < l((a',b]) + \epsilon$ and $\mathrm{cl}\left((a',b]\right) = [a',b] \subset (a,b]$.

Now let $\{A_n\}$ be a decreasing sequence of events in $\mathcal{G}$ such that $A_n \downarrow \emptyset$. We need to show that $\lim_{n\to\infty} l(A_n) = 0$; equivalently, for any $\epsilon > 0$, there exists N such that $l(A_n) < \epsilon$ when $n \ge N$. For each n, choose $B_n \in \mathcal{G}$ so that $\mathrm{cl}(B_n) \subset A_n$ and $l(A_n - B_n) < \epsilon/2^n$. Now $\cap_{n=1}^{\infty} \mathrm{cl}(B_n) \subset \cap_{n=1}^{\infty} A_n = \emptyset$. But the sets $\mathrm{cl}(B_n)$ are compact and therefore, there must be some $N < \infty$ such that $\cap_{k=1}^{N} B_k \subset \cap_{k=1}^{N} \mathrm{cl}(B_k) = \emptyset$. Hence $l(\cap_{k=1}^{N} B_k) = 0$. For $n \ge N$,

$$A_n - \bigcap_{k=1}^{N} B_k \subset \bigcup_{k=1}^{N} [A_n - B_k] \subset \bigcup_{k=1}^{N} [A_k - B_k].$$

Thus, $l(A_n) = l(A_n - \cap_{k=1}^{N} B_k) \le \sum_{k=1}^{N} l(A_k - B_k) < \epsilon$ for $n \ge N$, and the proof is complete. □

1.7 Independence

1.7.1 Definitions and simple examples

Suppose A and B are two events and $\mathbb{P}(B) > 0$. The ratio

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

is called the *conditional probability of A given B*. It is interpreted as the probability A occurs given on knows B has occurred, but nothing else. To see why, consider the frequentist interpretation of probability. In this view, $\mathbb{P}(B)$ is the frequency of occurrence of B and $\mathbb{P}(A \cap B)$ that of $A \cap B$ in repeated trials. Hence, $\mathbb{P}(A \mid B)$ is the frequency with which A occurs among those trials in which B does. If $\mathbb{P}(A \mid B)$ represents the probability of A knowing that B has occurred, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ means that occurrence of B does not affect our assessment of the probability of A. We formalize this in the notion of independence.

Definition 1.7.1 Two events A and B in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.4}$$

This definition is stated so as to be symmetric in A and B. Clearly, if A and B satisfy condition (1.4) and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B)/\mathbb{P}(B) = \mathbb{P}(A)$, and if $\mathbb{P}(A) > 0$, then $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. But the definition is stated to make sense when one or both of the probabilities can be zero. Two simple facts are worth noting immediately. First, if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, and B is any other event, then A and B are independent. Secondly, since $A \cap A = A$, an event is independent of itself if $\mathbb{P}(A) = \mathbb{P}^2(A)$, but this is true if and only if $\mathbb{P}(A) = 0$ or 1.

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two, different probability measures on $(\Omega, \mathcal{F})$. It may be that events A and B in $\mathcal{F}$ are independent when $\mathbb{P}_0$ is the probability measure, but not when $\mathbb{P}_1$ is. When clarity is needed, we will use the locution, "A and B are independent with respect to $\mathbb{P}$", to indicate the probability measure being used. When we just say, "A and B are independent", we have in mind an already fixed probability space.

Example 1.7.1 Consider the coin toss model of the previous section. Let i and j be any two, different positive integers. We have labelled the outcomes of tosses i and j by the random variable X_i and X_j. There are four possible outcomes of (X_i, X_j), each equally likely, and heads and tail are equally likely on each toss. Hence for any pair (ω_1, ω_2) of 0's and 1's,

$$\mathbb{P}(X_i = \omega_1, X_j = \omega_2) = \frac{1}{4} = \mathbb{P}(X_i = \omega_1)\mathbb{P}(X_j = \omega_2).$$

In words, every event concerning i is independent of every event concerning toss j.

Although this example is very basic, it illustrates a point. Here is an equivalent way to state the fact that every event concerning toss i is independent of every event concerning toss j: every event in $\sigma(X_i)$ is independent of every event in $\sigma(X_j)$. This is perhaps too ornate a formulation for such a simple example. But generally, independence in probability is about independence between σ-algebras. Indeed, let's go back to the simple situation in which A and B are independent events. The smallest σ-algebra containing A is $\mathcal{A} = \{ A, A^{\mathcal{C}}, \Omega, \emptyset \}$ and the smallest σ-algebra containing B is $\mathcal{B} = \{ B, B^{\mathcal{C}}, \Omega, \emptyset \}$.

Theorem 1.7.2 Events A and B are independent if and only if $\mathcal{A}$ is independent of $\mathcal{B}$.

Proof. The "if" direction is immediate. For the "only if" direction, since Ω and $\emptyset$ are independent of any event, it suffices to prove that if A and B are independent, then so A and $B^{\mathcal{C}}$ and also $A^{\mathcal{C}}$ and B. The second case differs from the first only by an interchange of the roles of A and B, so we will prove only the first one. To this end, observe that $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathcal{C}}) = \mathbb{P}(A)\mathbb{P}(B) + \mathbb{P}(A \cap B^{\mathcal{C}})$. Therefore,

$$\mathbb{P}(A \cap B^{\mathcal{C}}) = \mathbb{P}[1 - \mathbb{P}(B)] = \mathbb{P}(A)\mathbb{P}(B^{\mathcal{C}})$$

as was to be proved. □

These observations lead us to make the following definition:

Definition 1.7.2 Let $\mathcal{A}$ and $\mathcal{B}$ be families of events in a probability space. Then $\mathcal{A}$ and $\mathcal{B}$ are said to be independent if A and B are independent for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

We next move to a consideration of more than just two events.

Definition 1.7.3 Let $\mathcal{E}$ be a set of events. The events of $\mathcal{E}$ are (mutually) independent if for any m and distinct $B_1, ..., B_m \in \mathcal{E}$,

$$\mathbb{P}(B_1 \cap \cdots \cap B_m) = \mathbb{P}(B_1) \cdots \mathbb{P}(B_m)$$

Similarly, let $\{\mathcal{A}_\alpha \in \mathcal{I}\}$ be an arbitrary collection of families of events. The families of this collection are mutually independent if for any finite subset of distinct indices $\{\alpha_1, ..., \alpha_n\}$ from $\mathcal{I}$, and any events $A_1 \in \mathcal{A}_{\alpha_1}$, $A_2 \in \mathcal{A}_{\alpha_2}$, ..., $A_n \in \mathcal{A}_{\alpha_n}$, A_1, ..., A_n are independent.

For example, A_1, A_2, A_3 are mutually independent if they are pairwise independent, and, in addition, if $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$. It is important to remark on this last condition. Independence of A_1, A_2, A_3 is **not** equivalent to pairwise independence.

To simplify terminology the adjective 'mutually' will usually dropped. If more than two events or two families are said to be independent, the default assumption is that mutual independence is meant. When a collection of events is only pairwise independent, we will say so explicitly.

Theorem 1.7.2 generalizes. For example, the events $\{ A_1, ..., A_n \}$ are independent if and only if either generated σ-algebras, $A_i = \{ A_i, A_i^{\mathcal{C}}, \emptyset, \Omega \}$, $1 \leq i \leq n$ are independent. It follows that the same thing is true if we replace $\{ A_1, ..., A_n \}$ by any family of events.

1.7.2 Tossing a loaded coin

Using the definition of independence, we can generalize our model of an infinite number of tosses of a fair coin to a loaded coin. The two assumptions underlying the model will be: the probability of heads on any single toss is p, $p \neq \frac{1}{2}$, and that tosses are independent. Of course $0 < p < 1$. We shall work on the outcome space $\{ 0, 1 \}^\infty$ and use $X_i(\omega) = \omega_i$ to denote the outcome of toss i. For expressing probabilities conveniently, note that the formula $p^x(1-p)^x$ gives the probability p of heads if $x = 1$, and the probability $1 - p$ of tails if $x = 0$.

Again, proceed by first defining a finitely additive measure, P^p, on cylinder sets and then using Carathéodory's extension theorem. For the definition on cylinder sets, it is enough to define P^p on events of the form $\{\omega \in \{ 0, 1 \}^\infty : (\omega_1, ..., \omega_N) = (x_1, ..., x_N)\}$, where $(x_1, ..., x_N)$ is any specific sequence of 0's and 1's. Any cylinder set is a finite union of such events and so P^p is then extended to an arbitrary cylinder set by the requirement of finite additivity. But independence demands,

$$\begin{aligned} & P^p\left(\{\omega \in \{ 0, 1 \}^\infty; (\omega_1, \omega_2, ..., \omega_N) = (x_1, ..., x_N)\}\right) \\ =& \prod_{i=1}^{N} P^p\left(\{\omega; \omega_i = x_i\}\right) \\ =& \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i}. \end{aligned}$$

Theorem 1.6.6 implies that P^p extends to a countable additive probability measure $\mathbb{P}^p$ on $(\{ 0, 1 \}^\infty, \sigma(\mathcal{C}))$. Of course, this construction yields the fair-coin-toss space if $p = \frac{1}{2}$.

1.7.3 Percolation Models

The coin toss probability spaces have other applications. Here is one that has been the focus of considerable study in recent years. Consider an infinite graph with countably infinite vertex and edge sets. If the vertices represent points in a porous material, we can imagine that each edge is either open, meaning liquid can flow through it from on vertex to another, or closed, meaning no liquid can flow. Assume now that each edge is open independently with probability p. If we index the edges by the integers and imagine flipping coins independently to determine if edge i is open or closed, then we see that the infinite coin toss model also serves as a probability space for percolation. Of course, the questions one are much different. Percolation theory is mainly concerned about the structure, as a sub-graph, of the set of open edges.

1.8 Product Space and Independence

The fact that the different coin tosses are independent in the models constructed previously is intimately related to the fact that theses spaces are actually *product spaces*. In this section, we define a product space in general and show how it is related to independence. We will consider infinite product spaces only. The specialization of finite product spaces is easy. Also, we will omit the proof; they require new measure theory technique and the are not needed for the subsequent probability theory we want to cover.

Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $1 \leq i \leq \infty$, be probability spaces, each modelling the outcome of a different random experiment. We would like to create one large probability space that models all experiments at once, in such a way that their outcomes are independent from one another. This is a generalization of the problem of creating a probability space for infinite, independent sequence of coin tosses.

The outcome space will be the set of all sequences $\omega = \{\omega_1, \omega_2, ...\}$, where $\omega_i \in \Omega_i$, for each i. This is a product space which we denote by:

$$\bigotimes_{i=1}^{\infty} \Omega_i$$

A *'rectangle'* in this product space is a subset of the form:

$$B_1 \times B_2 \times \cdots \times B_N \times \Omega_{N+1} \times \Omega_{N+2} \times \cdots, \text{ where } B_i \in \mathcal{F}_i \text{ for } i \leq N,$$

and N is finite, positive integer. Let $\mathcal{R}$ be the set of all finite disjoint unions of such rectangles. Since the complement of a rectangles is a finite disjoint union of rectangles, as the reader may show, and since the intersection of two rectangle is a rectangle, $\mathcal{R}$ is an algebra. Define the product σ-algebra to be:

$$\bigotimes_{i=1}^{N} \mathcal{F}_i := \sigma(\mathcal{R}).$$

In the product space, the σ-algebra of events defined by the outcome of experiment i is the collection of events of the form:

$$\{\, (\omega_1, \omega_2, ...) \in \bigotimes_{i=1}^{\infty} \Omega_i; \omega_i \in B \,\}, \text{ where } B \in \mathcal{F}_i.$$

Let us denote this by $\widetilde{\mathcal{F}_i}$.

We are looking for a probability measure, $\mathbb{P}$, on this product σ-algebra so that the outcomes of the different experiments are independent and so that the probabilities of the outcomes for experiment i are given by $\mathbb{P}_i$, for each i. Mathematically, this means that $\mathbb{P}$ should satisfy:

$$\mathbb{P}(A) = \mathbb{P}_1(B_1)\mathbb{P}_2(B_2) \times \cdots \times \mathbb{P}_N(B_N) \tag{1.5}$$

on any rectangle A of the form:

$$B_1 \times B_2 \times \cdots \times B_N \times \Omega_{N+1} \times \cdots .$$

Theorem 1.8.1 There is a unique, countably additive measure, on $(\otimes_{i=1}^{\infty}\Omega_i, \otimes_{i=1}^{N}\mathcal{F}_i)$ satisfying (1.5). This measure is denoted by $\mathbb{P}_1 \times \mathbb{P}_2 \times \cdots$, and is called *product measure*. Under this measure, $\widetilde{\mathcal{F}_1}$, $\widetilde{\mathcal{F}_2}$,... are independent.

The strategy of the proof is the same as that we have used for coin tossing. First, construct a probability measure P on $\mathcal{R}$ by demanding that P satisfy (1.5) on rectangles and then using finite additivity to extend the definition to all of $\mathcal{R}$. Then show that P is continuous from above at the empty set. It's the second part of this argument that is harder. We have not imposed any metric space or topological structure and so we cannot use *Tychonoff's theorem*. It is necessary in this case to use some integration theory. This part of the proof is omitted.

1.9 Zero-One Law

We know that if A is an event independent of itself, then either $\mathbb{P}(A) = 0$ or 1. An important and interesting class of events called *tail events* have this property in infinite product spaces, a result which known as a *zero-one law*. In this section, we state zero-one laws for coin tossing. It is really just special case of a more general result. But the underlying idea is revealed most clearly in this special case, and so we defer discussing the general theory, which anyway is best done using random variables.

Let's remind ourselves the setting of coin tossing. We have $(\{\,0,\,1\,\}^{\infty}, \mathcal{B}(\{\,0,\,1\,\}^{\infty}), \mathbb{P}_p^{\infty})$ as outcome space for an infinite sequence of independent tosses of a coin, where p is the probability of heads. Let $\mathcal{C}_n$ be the σ-algebra of all events defined by the first n tosses, and let $\mathcal{C}^n$ be the σ-algebra of events that involve what happens in all tosses n, $n+1$, $n+2$, ... Also, recall that $X_i(\omega) = \omega_i$ denotes the outcome of toss i. It is helpful to note that $\mathcal{C}_n$ is the smallest σ-algebra with respect to which X_1, X_2, X_3, ..., X_n are all measurable, and $\mathcal{C}^n$ is the smallest σ-algebra with respect to which X_n, X_{n+1}, ... are all measurable. Obviously, the sequence $\{\mathcal{C}^n\}$ is decreasing. Let

$$\mathcal{C}^{\infty} := \bigcap_{n=1}^{\infty} \mathcal{C}^n.$$

This is called the *tail σ-algebra* for the coin tossing space and events in the tail σ-algebra are called *tail events*. Any event concerning the asymptotic behaviour of a sequence of tosses will be a tail event.

Lemma 1.9.1 Suppose that $(A_1, A_2, ...)$ is a sequence of events,

(i) If the sequence is increasing then $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$ is a tail event of the sequence;
(ii) If the sequence is decreasing then $\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$ is a tail event of the sequence.

Proof. If the sequence is increasing, then

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=k}^{\infty} A_n \in \sigma(A_k, A_{k+1}, ...)$$

for every $k \in \mathbb{N}_+$. If the sequence is decreasing then

$$\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=k}^{\infty} A_k \in \sigma\{A_k, A_{k+1}, ...\}$$

for every $k \in \mathbb{N}_+$. □

Corollary 1.9.2 Suppose that $(A_1, A_2, ...)$ is a sequence of events. Each of the following is a tail event of the sequence:

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i\,, \quad \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i.$$

Proof. The events $\bigcup_{i=n}^{\infty} A_i$ are decreasing in n and hence $\limsup_{n\to\infty} A_n = \lim_{n\to\infty} \bigcup_{i=n}^{\infty} A_i$ is a tail event by above lemma. The events $\bigcap_{i=n}^{\infty} A_i$ are increasing in n and hence $\liminf_{n\to\infty} A_n = \lim_{n\to\infty} \bigcap_{i=n}^{\infty} A_i$ is a tail event, again by the above lemma. □

Remark 1.9.3 We can deduce that, for any a, the event $\{\omega; \limsup \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \geq a\}$ and $\{\omega; \liminf \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \geq a\}$ are tail events by setting $A_k = \{\omega; \frac{1}{k}\sum_{i=1}^{k} X_i \geq a\}$.

Theorem 1.9.4 The event $\{X_n \text{ converges as } n \to \infty\}$ is a tail event for the sequence.

Proof. The *Cauchy criterion* for convergence states that X_n converges as $n \to \infty$ if and only if for every $\epsilon > 0$ there exists $N \in \mathbb{N}_+$ such that if m, $n \geq N$, then $|X_n - X_m| < \epsilon$. In this criterion, we can without loss of generality take ϵ to be rational, and for a given $k \in \mathbb{N}_+$, we can insist that m, $n \geq k$. With these restrictions, the Cauchy criterion is a countable intersection of events, each of which is in the tail σ-algebra. □

We have shown that, for any n, $\mathcal{C}^{n+1}$ and $\mathcal{C}_n$ are independent σ-algebra under the measure $\mathbb{P}_p^{\infty}$. It follows that $\mathcal{C}^{\infty}$ and $\mathcal{C}_n$ are independent for any n. Therefore, the algebra $\cup_{n=1}^{\infty}\mathcal{C}_n$ and $\mathcal{C}^{\infty}$ are independent. Now $\cup_{n=1}^{\infty}\mathcal{C}_n$ generates $\mathcal{B}(\{0, 1\}^{\infty})$, by construction, hence, Theorem 1.10.5 implies $\mathcal{B}(\{0, 1\}^{\infty})$ and $\mathcal{C}^{\infty}$ are independent. But $\mathcal{C}^{\infty}$ is a sub-σ-algebra of $\mathcal{B}(\{0, 1\}^{\infty})$, and thus $\mathcal{C}^{\infty}$ is independent of itself! If an event is A independent of itself then $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}^2(A)$, and this can only happen if $\mathbb{P}(A) = 0$ or 1. The following theorem is an immediate consequence.

Theorem 1.9.5 Every event in the tail σ-algebra, $\mathcal{C}^\infty$ has either probability zero or probability one.

The law of large numbers (which we have not proved yet), says that with probability one, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = p$. Theorem 1.9.5 is not enough to prove this, but despite its simplicity, it gets close. To explain, let $U(\omega) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i(\omega)$. Clearly, $0 \le U(\omega) \le 1$ for all ω. Let $g(a) = \mathbb{P}_p^\infty(U \ge a)$. As verified above, $\{U \ge a\}$ is a tail event. Hence $g(a) = 0$ or $g(a) = 1$ for each a. Since the events $\{U \ge a\}$ decrease as a increases, $g(a)$ is a decreasing function of a. But, since $0 \le U(\omega) \le 1$ for all ω, then $g(0) = 1$ and $g(a) = 0$ for $a > 1$. Therefore, there exists a unique a^* in $[0,1]$ such that $\mathbb{P}_p^\infty(U(\omega) > a) = 0$ for $a > a^*$ and $\mathbb{P}(U(\omega) \ge a) = 1$ for $a < a^*$, and hence

$$\mathbb{P}_p^\infty\left(\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \omega_i = a^*\right) = 1$$

This is already an interesting result. The same reasoning applied to the limit infinimum shows that there is a unique $a_* \in [0,1]$ such that

$$\mathbb{P}_p^\infty\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \omega_i = a_*\right) = 1$$

If we could prove that $a_* = a^*$, it would follow that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \omega_i$ exists with probability one. Unfortunately, there is no way to do this, as far as I know, using just the zero-one law. Of course, we will later deduce from the law of large numbers that

$$\mathbb{P}_p^\infty\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \omega_i = p\right) = 1$$

This result requires 'hard' analysis.

1.10 Appendix for Chapter I

1.10.1 The monotone class theorem

It is often necessary in probability theory to prove that all the events in a σ-algebra have a certain property. But there is rarely a simple characterization of the σ-algebra that allow the property to be checked directly. This section discusses indirect approaches that are used frequently.

Definition 1.10.1 Let $\mathcal{G}$ be a non-empty collection of subsets of Ω. $\mathcal{G}$ is called a *monotone class* if it is closed under countable increasing unions and countable decreasing intersections.

Let $\mathcal{S}$ be any non-empty collection of subsets of Ω. The power set of Ω is of course a monotone class. Thus, there is at least one monotone class containing $\mathcal{S}$. Also, it follows directly from the definition that the intersection of any family of monotone classes is a monotone class. Therefore,

$$M(\mathcal{S}) := \bigcap \{ \mathcal{M};\, \mathcal{S} \subset \mathcal{M},\, \mathcal{M} \text{ is a monotone class.} \}$$

is well defined and is the smallest monotone class containing $\mathcal{S}$. It is called the *monotone class generated by* $\mathcal{S}$. Note that any σ-algebra is a monotone class, and hence $M(\mathcal{S}) \subset \sigma(\mathcal{S})$. The next theorem states a necessary condition for the reverse inclusion.

Theorem 1.10.1 *(Monotone Class Theorem)* Let $\mathcal{R}$ be an algebra. Then $M(\mathcal{R}) = \sigma(\mathcal{R})$.

The proof will be given shortly. Instead, we now show how the theorem is used. The first result implies the uniqueness claim made in Carathéodory's extension theorem.

Theorem 1.10.2 Let $\mathcal{C}$ be an algebra of subsets of Ω. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be probability measures on $(\Omega, \sigma(\mathcal{C}))$ such that $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for every $A \in \mathcal{C}$. Then $\mathbb{P}_1 = \mathbb{P}_2$.

Proof. Let $\mathcal{M} = \{ A \in \sigma(\mathcal{C});\, \mathbb{P}_1(A) = \mathbb{P}_2(A) \}$. Then $\mathcal{C} \subset \mathcal{M}$ by assumption. Also, by the continuity properties of probability measure $\mathcal{M}$ is a monotone class. To see why, if $A_1 \supset A_2 \supset A_3 \supset \cdots$ is a decreasing sequence of sets in $\mathcal{M}$. Then, using continuity from above of probability measures, $\mathbb{P}_1(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}_1(A_n) = \lim_{n\to\infty} \mathbb{P}_2(A) = \mathbb{P}_2(\cap_{n=1}^{\infty} A_n)$. Therefore, $\cap_{n=1}^{\infty} A_n \in \mathcal{M}$. A similar calculation, using continuity from below, shows that $\mathcal{M}$ is closed under countable increasing unions. Since $\mathcal{M}$ is a monotone class containing $\mathcal{C}$, the monotone class theorem implies, $\sigma(\mathcal{C}) \subset \mathcal{M}$, and thus $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \sigma(\mathcal{C})$. □

The next theorem also implies the uniqueness of extension, but is a much stronger statement. It is an approximation argument, which says that all sets in σ-algebra generated by algebra can be approximated well by the sets in algebra.

Theorem 1.10.3 Let $\mathbb{P}$ be a probability measure on $\sigma(\mathcal{C})$ where $\mathcal{C}$ is an algebra. Then for every $\epsilon > 0$ and for every set $A \in \sigma(\mathcal{C})$, there is a set $A_0 \in \mathcal{C}$ such that $\mathbb{P}(A \Delta A_0) < \epsilon$.

Proof. Let $\mathcal{M}$ be the set of all $A \in \sigma(\mathcal{C})$ such that for every $\epsilon > 0$ there is a set $A_0 \in \mathcal{C}$ such that $\mathbb{P}(A \Delta A_0) < \epsilon$. Clearly, $\mathcal{C} \subset \mathcal{M}$. The proof will be done if we show $\mathcal{M}$ is a monotone class.

First observe that $\mathcal{M}$ is closed under taking complements, because $[A \Delta B]^{\mathcal{C}} = A^{\mathcal{C}} \Delta B^{\mathcal{C}}$. Now let $A_n \uparrow A$ and assume $A_n \in \mathcal{M}$ for each n. Fix $\epsilon > 0$ and choose k so large that $\mathbb{P}(A - A_k) < \frac{\epsilon}{2}$. Since $A_k \in \mathcal{M}$, there exists $A_{k0} \in \mathcal{C}$ such that $\mathbb{P}(A_k \Delta A_{k0}) < \frac{\epsilon}{2}$. Because $A_k \subset A$, $A \Delta A_{k0} \subset [A - A_k] \cup [A_k \Delta A_{k0}]$. Thus, $\mathbb{P}(A_k \Delta A_{k0}) < \epsilon$. This proves $A \in \mathcal{M}$, since $\epsilon > 0$ was arbitrary. Hence, $\mathcal{M}$ is closed under increasing unions. And, since $\mathcal{M}$ is closed under compliments, closure under increasing unions implies closure under decreasing intersections, as well. Thus $\mathcal{M}$ is indeed a monotone class. □

Let's prove the *monotone class theorem*. We know already $M(\mathcal{R}) \subset \sigma(\mathcal{R})$. Thus, we need only prove the reverse inclusion under the hypothesis of the theorem.

First note that any monotone class $\mathcal{G}$ which is also an algebra in fact a σ-algebra. Indeed, let $\mathcal{G}$ be both a monotone class and an algebra. Let $A_1, A_2, ...$ be a countable sequence of events in $\mathcal{G}$. then, for every n, $B_n := \cup_{j=1}^{n} A_j$ is an element of $\mathcal{G}$, by the definition of algebra. Since $\{B_n\}$ is an increasing sequence, $\cup_{n=1}^{\infty} B_n$ is also in $\mathcal{G}$ by the monotone class property of $\mathcal{G}$. Hence it follows that $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n$ is in $\mathcal{G}$. Thus $\mathcal{G}$ is closed under countable unions, making $\mathcal{G}$ a σ-algebra.

Therefore the proof will be complete if it is shown that $M(\mathcal{R})$ is an algebra. To this end, fix $A \in \mathcal{R}$ and consider the class of sets:

$$\mathcal{M}_A := \{\, B \,|\, B \subset \Omega,\, A \cup B \in M(\mathcal{R}) \,\}.$$

Clearly, $\mathcal{R} \subset \mathcal{M}_A$ since $\mathcal{R}$ is an algebra. Moreover, it is easy to check that $\mathcal{M}_A$ is a monotone class. Hence, $M(\mathcal{R}) \subset \mathcal{M}_A$ by definition of $M(\mathcal{R})$ as minimal. It follows that

$$A \cup B \in M(\mathcal{R}) \textit{ for every } A \in \mathcal{R} \text{ and every } B \in M(\mathcal{R}). \tag{1.6}$$

Now, let $E \in M(\mathcal{R})$ and consider $\mathcal{M}_E$. Observe that $\mathcal{R} \subset \mathcal{M}_E$ because of (1.6), and that, again, $\mathcal{M}_E$ is a monotone class. Hence, $M(\mathcal{R}) \subset \mathcal{M}_E$, which says that

$$E \cup B \in M(\mathcal{R}) \textit{ for every } E \in M(\mathcal{R}) \textit{ and every } B \in M(\mathcal{R}).$$

This proves that $M(\mathcal{R})$ is closed under finite unions. To show $M(\mathcal{R})$ is closed under complements we hos that $\mathcal{M} := \{A; A \subset \Omega, A^{\mathcal{C}} \in M(\mathcal{R})\}$ is a monotone class containing $\mathcal{R}$, and hence that $M(\mathcal{R}) \subset \mathcal{M}$.

1.10.2 Dynkin's Theorem

There is another theorem of monotone class type that is sometimes more convenient to use.

Definition 1.10.2 A class $\mathcal{J}$ of subsets of a set Ω is called a λ-system if (i) $\Omega \in \mathcal{J}$; (ii) $A, B \in \mathcal{J}$ and $A \subset B$ imply that $B - A \in \mathcal{J}$; and (iii) $\mathcal{J}$ is closed under countable increasing unions.

Let $\mathcal{S}$ be a non-empty collection of subsets of Ω. The intersection of all λ-system containing $\mathcal{S}$ is a λ-system. It will be denoted by $\mathcal{J}(\mathcal{S})$. Of course, a σ-algebra is a λ-system, and hence $\mathcal{J}(\mathcal{S}) \subset \sigma(\mathcal{S})$. We know from the monotone class theorem that the reverse inclusion is true if $\mathcal{S}$ is an algebra. The next result says that a much weaker condition on $\mathcal{S}$ actually suffices.

Theorem 1.10.4 *(Dynkin's $\pi - \lambda$ system theorem)* Let $\mathcal{C}$ be a family of subsets of Ω closed under finite intersection. Then $\sigma(\mathcal{C}) = \mathcal{J}(\mathcal{C})$.

This theorem is called the $\pi - \lambda$ system theorem because sometimes a family closed under finite intersection is called a π-system. The advantage of this theorem over the monotone class theorem is the simpler requirement on the generating collection, $\mathcal{C}$. The Dynkin's system theorem gives a handy criterion for checking when two σ-algebras are independent.

Theorem 1.10.5 Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be closed under finite intersection. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent, then so are $\sigma(\mathcal{G}_1)$ and $\sigma(\mathcal{G}_2)$.

Proof. Let $A \in \mathcal{G}_1$. Let $\mathcal{E}_A$ be the set of all events B which is independent of A. By assumption, $\mathcal{E}_A$ contains $\mathcal{G}_2$. Also, $\Omega \in \mathcal{E}_A$ because Ω is independent of all events, and if B_n is an increasing sequence in $\mathcal{E}_A$, and $B = \cup_{n=1}^{\infty} B_n$, then

$$\mathbb{P}(A \cap B) = \lim_{n\to\infty} \mathbb{P}(B_n \cap A) = \lim_{n\to\infty} \mathbb{P}(B_n)\mathbb{P}(A) = \mathbb{P}(B)\mathbb{P}(A).$$

Hence, $B \in \mathcal{E}_A$ also. Finally, if $B_1 \subset B_2$ and both sets are in $\mathcal{E}_A$,

$$\begin{aligned}\mathbb{P}(B_2)\mathbb{P}(A) &= \mathbb{P}(B_2 \cap A) = \mathbb{P}(B_1 \cap A) + \mathbb{P}((B_2 - B_1) \cap A) \\ &= \mathbb{P}(B_1)\mathbb{P}(A) + \mathbb{P}((B_2 - B_1) \cap A).\end{aligned}$$

By subtracting $\mathbb{P}(B_1)\mathbb{P}(A)$ from both sides,

$$\mathbb{P}((B_2 - B_1) \cap A) = \big[\mathbb{P}(B_2) - \mathbb{P}(B_1)\big]\mathbb{P}(A) = \mathbb{P}(B_2 - B_1)\mathbb{P}(A)$$

Thus $B_2 - B_1$ and A are independent, which implies $B_2 - B_1 \in \mathcal{E}_A$. We have shown that $\mathcal{E}_A$ is a λ-system. By the Dynkin system theorem, $\sigma(\mathcal{G}_2) \subset \mathcal{E}_A$. This argument works for any $A \in \mathcal{G}$, and therefore we have shown that A and B are independent whenever $A \in \mathcal{G}_1$ and $B \in \sigma(\mathcal{G}_2)$.

Now fix $B \in \sigma(\mathcal{G}_2)$ and let $\mathcal{G}_B = \{\, U;\, U \text{ and } B \text{ are independent}\}$. By repeating the argument above with minor modification, we can show that $\mathcal{G}_B$ is a λ-system. Since it contains $\mathcal{G}$, it follows that, A and B are independent for any $A \in \sigma(\mathcal{G}_1)$. Since the argument works for all $b \in \sigma(\mathcal{G}_2)$, we have shown $\sigma(\mathcal{G}_1)$ and $\sigma(\mathcal{G}_2)$ are independent. □

Example 1.10.6 Consider the coin toss space $(\{\,0,\,1\,\}^{\infty}, \mathcal{B}(\{\,0,\,1\}^{\infty}), \mathbb{P}_p^{\infty})$ defined previously. It should be intuitively clear, for example, that any event defined by what happens on even numbered tosses are independent of any event defined by what happens on odd numbered tosses. More generally, events generated by disjoint sets of tosses are independent. For the sake of notational simplicity, we will prove a special case of this claim. The reader may easily generalize the statement and proof, not just for coin toss space, but for any infinite product of probability spaces.

Recall that $\mathcal{C}_N$ denoted the σ-algebra of events defined by the first N tosses. Let $\mathcal{C}^{N+1}$ be the σ-algebra generated by the tosses that occur after toss N. This is smallest σ-algebra containing events of the form $\{\omega_i; \omega_i \in B\}$, where $i \geq N+1$, and $B \subset \{\,0,\,1\,\}$. We want to show that $\mathcal{C}_N$ and $\mathcal{C}^{N+1}$ are independent. By Dynkin's $\pi - \lambda$ system theorem, it suffices to

find π-system $\mathcal{G}$ and $\mathcal{H}$such that $\mathcal{G}$ and $\mathcal{H}$ are independent and $\mathcal{C}_N = \sigma(\mathcal{G})$ and $\mathcal{C}^{N+1} = \sigma(\mathcal{H})$. The following choices work. Let $\mathcal{G}$ by the collection of all events of the form:

$$U = \{\, \omega;\, \omega_i \in B_i,\, 1 \le i \le N \,\} = B_1 \times B_2 \times \cdots \times B_N \times \{\, 0,\, 1 \,\}^\infty \tag{1.7}$$

where $B_i \in \{\, 0,\, 1 \,\}$ for $1 \le i \le N$. $\mathcal{G}$ is just the collection of finitely-based 'rectangles' based on the first N coordinates of $\{\, 0,\, 1 \,\}^\infty$. $\mathcal{G}$ certainly includes the event that toss i is heads or that it is tails, for every i, $1 \le i \le N$, and so $\mathcal{G}$ generates $\mathcal{C}_N$. Similarly, let $\mathcal{H}$ be the collection of events of the form:

$$V = \{\, \omega;\, \omega_i \in B_i,\, N+1 \le i \le N+m \,\} = \{\, 0,\, 1 \,\}^N \times B_{N+1} \times \cdots \times B_{N+m} \times \{\, 0,\, 1 \,\}^\infty, \tag{1.8}$$

where $B_i \subset \{\, 0,\, 1 \,\}$, for $N+1 \le i \le N_m$, and m is a positive integer. This is the class of rectangles, based on a finite number of tosses occurring after toss N. Again, $\mathcal{H}$ generates $\mathcal{C}^{N+1}$ and $\mathcal{H}$ is closed under finite intersection.

Let $\mathbb{P}_1^p$ denote the measure on $\{\, 0,\, 1 \,\}$ when the probability of 1 is p. Now suppose that U and V are given in (1.7) and (1.8). Then, because, $\mathbb{P}^p$ is a product measure,

$$\begin{aligned}
&\mathbb{P}^p(U \cap V) = \mathbb{P}^p\left(B_1 \times \cdots \times B_N \times B_{N+1} \times \cdots \times B_{N_m} \times \{\, 0,\, 1 \,\}^\infty\right) \\
=&\prod_{i=1}^{n} \mathbb{P}_1^p(B_i) \cdot \prod_{i=N+1}^{N+m} \mathbb{P}_1^p(B_i) \\
&= \mathbb{P}^p(U)\mathbb{P}^p(V).
\end{aligned}$$

This calculation shows that $\mathcal{G}$ and $\mathcal{H}$ are independent, and it then follows from Theorem 1.10.5 that $\mathcal{C}_N$ and $\mathcal{C}^{N+1}$ are independent. This example is easily generalized to the infinite product space constructed in previous section.

Chapter 2

Random Variable and Expectation

2.1 Random Variables: Basics Definition

We have already defined the notion of random variable in *Chapter 1*, but not delving into the details. Let's do so by starting to repeat the definition:

Definition 2.1.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \mapsto \mathbb{R}$ is called a *random variable*, if $X^{-1}(U) \in \mathcal{F}$ whenever U is a Borel subset of $\mathbb{R}$.

In measure theory, a function $f : U \mapsto V$, between a space U with σ-algebra $\mathcal{G}$ and a space V with a σ-algebra $\mathcal{H}$, is called $\mathcal{G}/\mathcal{H}$-*measurable* if $f^{-1}(U) \in \mathcal{G}$ for every $U \in \mathcal{H}$. When V is $\mathbb{R}^n$ and $\mathcal{H}$ is the Borel σ-algebra of $\mathbb{R}^n$, a $\mathcal{G}/\mathcal{B}(\mathbb{R}^n)$-*measurable* function is said to be *Borel measurable*. Thus, a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ is a real-valued, Borel-measurable function on $(\Omega, \mathcal{F})$.

Proposition 2.1.1 Let $\mathcal{R}$ be any class of subsets of $\mathbb{R}$ which generates the Borel subsets; that is, $\sigma(\mathcal{R}) = \mathcal{B}(\mathbb{R})$. Then X is a random variable if and only if $X^{-1}(U) \in \mathcal{F}$ for all $U \in \mathcal{R}$.

The proof is straightforward, we leave to the readers to prove. Examples of simple families of subsets generating the Borel sets are intervals of the form (a, ∞), or all intervals of the form $(-\infty, a)$, etc. Before proceeding further, let's clarify the notation: it is common to write $X^{-1}(U) = \{\omega; X(\omega) \in U\}$ simply as $\{X \in U\}$ and to abbreviate $\mathbb{P}(\{X \in U\})$ by $\mathbb{P}(X \in U)$.

Definition 2.1.2 Define $\sigma(X) := \{\{X \in U\}; U \in \mathcal{B}\} = X^{-1}(\mathcal{B}(\mathbb{R}))$. This is called *the σ-algebra generated by X*. It is the σ-algebra of all events concerning the outcome of X. More generally, if $\{X_t; t \in \mathcal{T}\}$ is any collection of random variables on a common probability space $\sigma(\{X_t; t \in \mathcal{T}\})$ is the smallest σ-algebra containing $\sigma(X_t)$ for all $t \in \mathcal{T}$.

Example 2.1.2 Consider the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$, where λ is the Lebesgue measure. Let $X(\omega) = \omega$ for $0 \leq \omega \leq 1$. This random variable just returns the value of the point randomly and uniformly selected from $[0,1]$, and for this reason is called the *canonical random variable* for the uniform distribution on $[0,1]$. One can use X on $([0,1], \mathcal{B}([0,1]), \lambda)$ as the model for a uniformly chosen point from $[0,1]$, rather than $([0,1], \mathcal{B}([0,1]), \lambda)$ itself. The random variable formulation of the model is actually more convenient. Clearly, $\sigma(X) = \mathcal{B}([0,1])$.

Any Borel-measurable function from $[0,1]$ to $\mathbb{R}$ defines a random variable, and thus $([0,1], \mathcal{B}([0,1]), \lambda)$ can support collections of random variables with a rich structure. For example, let

$$Y_k(\omega) = x_k, \text{ for } k \geq 1, \text{ if } 0 \leq \omega \leq 1,$$

where $\omega = \sum_{n=1}^{\infty} \frac{x_n}{2^n}$ is the unique, non-repeating, decimal expansion of ω, and let $Y_k(1) = 1$ for all k. (We have effectively alredy encountered these random variables before.) Each Y_k only takes the values 0 or 1; moreover,

$$\lambda(\{\omega; Y_k(\omega) = 0\}) = \lambda(\{\omega; Y_k(\omega) = 1\}) = \frac{1}{2},$$

for each k, since $\{\omega; Y_k(\omega) = 0\}$ is the disjoint union

$$\bigcup_{j=0}^{2^k-1} [\frac{j}{2^{k-1}}, \frac{j}{2^{k-1}} + \frac{1}{2^k}).$$

Thus each Y_k may be regarded as a random variable model for the toss of a fair coin. Later, it will be seen that $\{Y_1, Y_2, ...\}$, on the space $(([0,1], \mathcal{B}([0,1]), \lambda))$, models a sequence of independent fair coin tosses.

If $(x_1, x_2, ..., x_n) \in \{0, 1\}^n$, the event $\{Y_1 = x_1, ..., Y_n = x_n\} = [a, a + 2^{-n})$ where a is the dyadic rational $\sum_{k=1}^{n} x_k 2^{-k}$. These sets generate the Borel sets of $[0,1]$. Thus, $\sigma(Y_1, Y_2, ...) = \mathcal{B}([0,1])$.

Example 2.1.3 Let $(\{0, 1\}^\infty, \mathcal{B}(\{0, 1\}^\infty), \mathbb{P}_p^\infty)$ be the coin toss space constructed in Chapter 1. Recall that it models an infinite sequence of independent tosses of a coin, each having probability p of heads. Points $\{0, 1\}^\infty$ are denoted $\omega = (\omega_1, \omega_2, ...)$. The function, $X_i(\omega) = \omega_i$, picks out the result of toss i from ω. Thus the sequence $(X_1, X_2, ...)$ re-express the model $(\{0, 1\}^\infty, \mathcal{B}(\{0, 1\}^\infty), \mathbb{P}_p^\infty)$ in terms of random variables.

Let $Y(\omega) := \sum_{n=1}^{\infty} 2^{-n} X_n(\omega)$. Then Y is a random variable taking values in $[0,1]$. This is a sequence of the fact proved in the next section that limits of random variables are random variables. If $p = \frac{1}{2}$, then Y on $(\{0, 1\}^\infty, \mathcal{B}(\{0, 1\}^\infty), \mathbb{P}_p^\infty)$ is uniformly distributed on $[0,1]$ in the sense that $\mathbb{P}_{1/2}^\infty(Y \in U) = \lambda(U)$ if U is a Borel subset of $[0,1]$. This is intuitively clear from above example and may be proved by showing that the probability that Y in any interval $[a,b)$, where $0 < a, b \leq 1$, a and b are dyadic rationals, is $b - a$.

The random variables defined in Example 2.1.5 have a meaning whatever probability measure is put on $([0,1],\mathcal{B}([0,1]))$; more generally, whether or not a real-valued function on $(\Omega,\mathcal{F})$ is a Borel-measurable has nothing to do with a measure. Nevertheless, it is always assumed that there is some fixed probability measure, $\mathbb{P}$, on the outcome space on which a random variable is defined, because this measure determines the *distribution* of the random variable, as defined next.

Definition 2.1.3 Let X be a *random variable* defined on $(\Omega,\mathcal{F},\mathbb{P})$. The *distribution measure* of X is the probability measure $\mathbb{F}_X := \mathbb{P}\circ X^{-1}$ on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$; that is

$$\mathbb{F}_x(U) = \mathbb{P}(\{X\in U\}),\ U\in\mathcal{B}(\mathbb{R}).$$

The *cumulative distribution function (cdf)* of X is

$$F_X(x) := \mathbb{P}(\{X\le x\}),\ -\infty < x < \infty.$$

Note that $F_X(x) = \mathbb{F}_X\left((-\infty,x]\right)$.

A random variable, X, is called *discrete* if it takes values in a countable subset $\{x_1,x_2,...\}$ of $\mathbb{R}$. In this case,

$$\mathbb{F}_X(U) = \sum_{i;x_i\in U}\mathbb{P}(X=x_i).$$

On the other hand, a random variable X is called *continuous*, if there exists a Borel function f_X called the *probability density function (pdf)* of X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds,\ \text{for all } x.$$

When the density exists, it is defined only up to Lebesgue almost-everywhere equivalence: if $g(s)=f(s)$ Lebesgue-almost-everywhere and f is a pdf for X, the so is g. Even so, it is common to always speak of the pdf of a random variable, as if it were unique. We may always choose a version f_X such that $f_X(s)\ge 0$ for all s, and so we impose non-negativity on density function as a matter of definition.

Remark 2.1.4 A continuous random variable is not necessarily continuous as a function of ω! Rather, its cdf is an absolutely continuous function of x. In applications, random variables are usually either discrete or continuous, but, as we shall se, there can be random variables whose cdf's are continuous but not continuous random variables.

Example 2.1.5 Let X be as defined in Example , since $X(x)$ for $0\le x\le 1$,

$$\mathbb{F}_X(U) = \lambda\left(\{x\in[0,1]; x = X(x)\in U\}\right) = \lambda\left(U\cap[0,1]\right).$$

The distribution function is thus:

$$F_X(x) = \lambda\left((-\infty,x]\cap[0,1]\right) = \begin{cases} 0, & \text{if } x<0;\\ x, & \text{if } 0\le x\le 1;\\ 1, & \text{if } x>1.\end{cases}$$

It follows that X is a continuous random variable with density

$$f_X(x) = \mathbf{1}_{(0,1)}(x).$$

Consider next the random variable $Y_1(x)$ defined in Example 2.1.5 again. This may be rewritten as $Y_1(x) = \mathbf{1}_{[\frac{1}{2},1]}(x)$, $0 \leq x \leq 1$. It is a discrete random variable and its distribution measure is

$$\mathbb{F}_{Y_1}(U) = \frac{1}{2}\left(\delta_0(U) + \delta_1(U)\right),$$

where δ_z denotes the *Dirac delta measure* at $z : \delta_z(U) = \mathbf{1}_U(z)$.

If $X_1, ..., X_n$ are random variables defined on the same probability space, we think of $Z = (X_1, ..., X_n)$ as a random vector. In this case, $Z : \Omega \mapsto \mathbb{R}^n$ is a Borel-measurable map. This will be proved in the next section; Meanwhile, assuming it is true makes the next definition possible.

Definition 2.1.4 Let $Z = (X_1, ..., X_n)$ be a random vector. The *distribution measure* of Z is the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ defined by:

$$\mathbb{F}_Z(U) = \mathbb{P}(Z \in U),\ U \in \mathcal{B}(\mathbb{R}^n),$$

and the *joint cumulative distribution function* of Z is

$$F_Z(x_1, ..., x_n) := \mathbb{P}\left(X_1 \leq x_1, ..., X_n \leq x_n\right),\ (x_1, ..., x_n) \in \mathbb{R}^n.$$

$X_1, ..., X_n$ are said to be *jointly continuous* if there is a Borel-measurable function, f_Z, which can be assumed to be non-negative and is called the *joint density* of $X_1, ..., X_n$, such that

$$F_Z(x_1, ..., x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_Z(s_1, ..., x_n)\, ds_1 \cdots ds_n.$$

A *stochastic process* is a collection of random variables defined on a common probability space. A random vector is thus formally an example of a stochastic process, but usually latter term is reserved for collections, $\{X_t; t \in \mathcal{T}\}$ indexed by an infinite set $\mathcal{T}$. The sequence $\{Y_1, Y_2, ...\}$ of coin toss random variables defined in Example 2.1.5 is a simple example o stochastic process. Given a stochastic process, $\{X_t; t \in \mathcal{T}\}$, and a finite subset $\{t_1, ..., t_n\}$ of distinct elements of $\mathcal{T}$, the distribution measure $\mathbb{F}_{(X_{t_1}, ..., X_{t_n})}$ is called a finite-dimensional distribution of the process. We shall see later that for $(Y_1, Y_2, ...)$ in Example 2.1.3,

$$\mathbb{F}_{(Y_{t_1}, ..., Y_{t_n})} = \mathbb{F} \otimes \cdots \otimes \mathbb{F} \ \ n \text{ times}$$

where $\mathbb{F}$ is the distribution measure for any single Y_k; namely, $\mathbb{F} = \frac{1}{2}(\delta_0 + \delta_1)$. This is equivalent to saying that $(Y_1, Y_2, ...)$ models independent tosses of fair coin.

Finally, many problems and models generate functions on probability space that are not real-valued. We have already seen one example, a random vector $(X_1, ..., X_n)$ with

values in $\mathbb{R}^n$. We could also interpret a stochastic process $\{X_t; t \in \mathcal{T}\}$ as a random map, $X_{\mathcal{T}}(\omega) = \{X_t(\omega); t \in \mathcal{T}\}$, from Ω to $\mathbb{R}^{\mathcal{T}}$, where $\mathbb{R}^{\mathcal{T}}$ is the product space consisting of all functions from $\mathcal{T}$ to $\mathbb{R}$. To formalize such extension of the concept of random variable requires imposing appropriate measurability conditions. We will deal with this only when needed, except for one case we treat now. It is useful when studying limits of random variables to allow them to take values in the extended real numbers, defined as $\bar{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{\infty\}$. The extended reals are an ordered set carrying the natural ordering. It is a compact topological space if open subset of $\mathbb{R}$ are still open and if the open neightborhoods of ∞ are taken to be sets of the form $(a, \infty) \cup \{\infty\}$, thus, for $-\infty$, $\{-\infty\} \cup (-\infty, a)$. It is easily checked that the Borel-σ-algebra, $\mathcal{B}(\bar{\mathbb{R}})$, is the collection of all sets U such that either U is a Borel subset of $\mathbb{R}$, or U has one of the form $V \cup \{\infty\}$, $V \cup \{-\infty\}$, or $V \cup \{-\infty, \infty\}$, where V is a Borel set of $\mathbb{R}$.

Definition 2.1.5 An extended random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is an $\mathcal{F}/\mathcal{B}(\bar{\mathbb{R}})$-measurable map from Ω to $\bar{\mathbb{R}}$.

It is useful to note that Proposition 2.1.1 is true for extended random variables. Merely note, for instance, that $\{X = \infty\} = \bigcap_{n=1}^{\infty} \{X > n\}$ and, hence, $\{X = \infty\} \in \mathcal{F}$ if $\{X > a\}$ is for all a. The reader should be able to complete the argument with this hint.

2.2 Basic facts about random variables

The first theorem is about the preservation of measurability under various operations on random variables. The point is that standard operations on random variables result again in random variables. We use these results all the time, almost without thinking about them. The probability measure actually plays no role in this theorem, and the results are purely about measurable functions on a measure space. The proofs of the different claims are elementary, but not that important for our later work.

Theorem 2.2.1 (i) Let $X_1, X_2, ..., X_n$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $Z(\omega) = (X_1(\omega), ..., X_n(\omega))$ defines a Borel-measurable map from Ω to $\mathbb{R}^n$. Conversely, if $Z(\omega) = (X_1(\omega), ..., X_n(\omega))$ is Borel-measurable, each X_k is a random variable;

(ii) Let $X_1, X_2, ..., X_n$ be random variables on a common probability space. Then for any function $h : \mathbb{R}^n \mapsto \mathbb{R}$ which is $\mathcal{B}(\mathbb{R}^n)/\mathcal{B}(\mathbb{R})$-measurable, $h(X_1, ..., X_n)$ is a random variable. In particular, any algebraic function of random variable is a random variable;

(iii) If X and Y are random variables on a common probability space, so is any linear combination $aX + bY$;

(iv) If $\{X_n\}$ is a sequence of random variables on a common probability space, $\sup X_n$, $\inf X_n$, $\limsup X_n$ and $\liminf X_n$ are extended random variables. Also,

$$X(\omega) := \begin{cases} \lim X_n(\omega), & \text{if } \liminf X_n(\omega) = \limsup X_n(\omega); \\ 0, & \text{otherwise.} \end{cases}$$

is a random variable.

Proof. (i) It is necessary to show that $Z^{-1}(U) \in \mathcal{F}$ for each Borel subsets U contained in $\mathbb{R}^n$. For this, it suffices to show $Z^{-1}(V) \in \mathcal{F}$ for every V of the form $V = B_1 \times \cdots \times B_n$, where B_i, $1 \leq i \leq n$, are Borel subsets of $\mathbb{R}$, since sets of this form generate the Borel-σ-algebra of $\mathbb{R}^n$. But

$$Z^{-1}\left(B_1 \times \cdots \times B_n\right) = \bigcap_{i=1}^{n} \{X_i \in B_i\},$$

and this belongs to $\mathcal{F}$ because each set $\{X_i \in B_i\}$, $1 \leq i \leq n$ does.

(ii) Let $Y = h(X_1, ..., X_n) = h(Z)$. Then if U is a Borel subset of $\mathbb{R}$,

$$\{Y \in U\} = Z^{-1}(h^{-1}(U)).$$

Since h is Borel-measurable, $h^{-1}(U)$ is a Borel-measurable subsets of $\mathbb{R}^n$. By part (i), $Z^{-1}\left(h^{-1}(U)\right) \in \mathcal{F}$. (iii) is a direct result following from (ii). We leave (iv) to the readers to prove. □

The next theorem summarizes the important properties of cumulative distribution functions and how they relate to distribution measures and probability density functions. In this theorem, $F(x-)$ is notation for the left-limit, $\lim_{s \uparrow x} F(s)$.

Theorem 2.2.2 Let $\mathbb{F}_X$ and F_X be, respectively, the distribution measure and the cdf of a random variable X.

(i) F_X is an *increasing, right-continuous function* satisfying $\lim_{s \downarrow -\infty} F_X(x) = 0$ and $\lim_{s \uparrow \infty} F_X(s) = 1$;
(ii) If f_X is a pdf for a random variable X, then $f_X \geq 0$ and $\int_{-\infty}^{\infty} f_X(s)ds = 1$. At points x where $f_X(x)$ is continuous, $F_X'(x) = f_X(x)$. In fact, $f_X(x) = F'(x)$ for Lebesgue-almost-every x;
(iii) For any x, $\mathbb{P}(X = x) = F_X(x) - F_X(x-)$. In particular, if F_X is continuous, $\mathbb{P}(X = x) = 0$ for all x;
(iv) The cdf uniquely determines the distribution measure: if $F_X = F_Y$, then $\mathbb{F}_X = \mathbb{F}_Y$. In particular, if X admits a pdf f_X,

$$\mathbb{F}_X(U) = \int_U f_X(s)ds, \text{ for any Borel set } U \tag{2.1}$$

The facts stated in this theorem generalize to random vectors.

Theorem 2.2.3 Let $Z = (X_1, ..., X_n)$ be a random vector.

(i) The joint cumulative distribution function F_Z is *right-continuous* in each variable and $\lim_{x_1,...,x_n \downarrow -\infty} F_Z(x_1, ..., x_n) = 0$ and $\lim_{x_1,...,x_n \uparrow \infty} F_Z(x_1, ..., x_n) = 1$;
(ii) If Z admits a joint density f_Z such that $f_Z \geq 0$ and $\int_{\mathbb{R}^n} f_Z(x_1, ..., x_n)dx_1...dx_n = 1$;

(iii) The joint cdf uniquely determines the distribution measure. If Z admits a joint density f_Z, then

$$\mathbb{P}(Z \in V) = \int_V f_Z(x_1, ..., x_n) dx_1 ... d_x n.$$

Proof. (Theorem 2.2.2) (i) It is clear that $F_X(x) = \mathbb{P}(X \leq x)$ is increasing in x. Let $x_n \downarrow x$. Since $\{X \leq x\} = \bigcap \{X \leq x_n\}$ is the intersection of a decreasing sequence of sets, the continuity of probability measure implies,

$$F_X(x) = \mathbb{P}(X \leq x) = \lim_{n\to\infty} \mathbb{P}(X \leq x_n) = \lim_{n\to\infty} F_X(x_n).$$

This proves right continuity. If $x_n \downarrow -\infty$, $\bigcap \{X_n \leq x\} = \emptyset$. Hence, $\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} \mathbb{P}(X \leq x_n) = 0$. If $x_n \uparrow \infty$, $\lim_{n\to\infty} F_X(x_n) = \mathbb{P}(X < \infty) = 1$.

(ii) It is assumed that $f_X(x) \geq 0$ by convention. By definition of the pdf, $\int_{-\infty}^{\infty} f_X(s) ds = \mathbb{P}(-\infty < X < \infty) = 1$. That $F_X'(x) = f_X(s)$ for Lebesgue almost every where x is the Lebesgue Differentiation Theorem.

(iii) If $F_X = F_Y$, then

$$\mathbb{F}_X((a, b]) = F_X(b) - F_X(a) = F_Y(b) - F_Y(a) = \mathbb{F}_Y((a, b]),$$

for all $a < b$. Thus $\mathbb{F}_X$ and $\mathbb{F}_Y$ agree on algebra of finite disjoint unions of intervals of the form $(a, b]$. This algebra generates the Borel σ-algebra, and so the assertion follows. Finally, consider equation (2.1) when X admits a pdf, f_X. The right-hand side defined a countably additive measure on Borel sets U. But

$$\mathbb{F}_X((a, b]) = F_X(b) - F_X(a) = \int_{(a,b]} f_X(s) ds$$

for any $a < b$, so, again, the two measure must agree on all Borel sets. □

Proof. (Theorem 2.2.3) The proof of parts (i) and (ii) are simple extensions of proofs of the corresponding statements of Theorem 2.2.2. By the same type of reasoning as in the proof of part (iv) of Theorem 2.2.2, to prove (iii) it suffice to show that the cumulative distribution function F_Z determines $\mathbb{F}_Z(\mathbb{R})$ for any rectangle in $\mathbb{R}^n$ of the form,

$$R = (a_1, b_1] \times \cdots \times (a_n, b_n].$$

This is left as an exercise. □

If $\mathbb{P}(X = Y) = 1$, we say that $X = Y$ $\mathbb{P}$-almost surely and we write $X \overset{a.s.}{=} Y$. We also say that X and Y are version of one another. Clearly, if Y is a version of X, then $F_X = F_Y$ and $\mathbb{P}(\{X \in U\} \Delta \{Y \in U\}) = 0$ for any Borel U. From the point of view of probabilistic

analysis, X and Y are the same random variable. For this reason, we should perhaps define a random variable as an equivalence class of almost surely equal, Borel-measurable functions. but this is to enter the realm of advanced pedantry. Rather, a random variable in this text is always a specific function on a probability space. When there is ambiguity in the definition of a random variable due to the possibility of different versions, we shall (usually!) point this out.

2.3 Random variables models

2.3.1 Random variables and vectors

A real-valued function f defined on $\mathbb{R}$ is called a probability density if it is non-negative and Borel-measurable, and if $\int_{\mathbb{R}} f(x)\,dx = 1$. It is natural to ask whether there is a random variable X such that $f_X = f$. If f is a probability density, then

$$\mathbb{F}(U) := \int \mathbf{1}_U(x) f(x)\,dx$$

defines a probability measure on the Borel sets of $\mathbb{R}$, so an equivalent question is whether there exists X such that $\mathbb{F}_X = \mathbb{F}$. More generally, if $\mathbb{F}$ should instead denote an arbitrary probability measure on $\mathbb{R}^n$, is there a random vector Z such that $\mathbb{F}_Z = \mathbb{F}$? The answer is yes and the construction of Z is simple. As the underlying probability space, take $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{F})$ itself. Let Z be the identity function, $Z(x_1, ..., x_n) = (x_1, ..., x_n)$. Then

$$\begin{aligned}\mathbb{F}_Z(U) &= \mathbb{F}\left(\{(x_1, ..., x_n); Z(x_1, ..., x_n) \in U\}\right) \\ &= \mathbb{F}\left(\{(x_1, ..., x_n); (x_1, ..., x_n) \in U\}\right) = \mathbb{F}(U)\end{aligned}$$

This is called the *canonical random variable* with distribution measure $\mathbb{F}$.

Suppose now that $F : \mathbb{R} \mapsto [0, 1]$ satisfies:

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

Theses properties are shared by the cumulative distribution function of any random variable. Is there then necessarily a random variable X for which $F_X = F$? The answer is again yes, and for this reason, any F satisfying above conditions is called cumulative distribution function.

One way to prove X exists is to show there is unique probability measure $\mathbb{F}$ on $\mathbb{R}$ such that $\mathbb{F}\left((a, b]\right) = F(b) - F(a)$ and then to define X to be the canonical random variable for

the distribution measure $\mathbb{F}$. The argument constructing $\mathbb{F}$ is a straightforward extension of the one we used in Chapter 1 to construct Lebesgue measure on $[0,1]$, and $\mathbb{F}$ is called the *Lebesgue-Stieltjes measure associated to F*. However, there is a slicker construction, which has practical applications to simulation and theoretical applications to coupling of random variables.

Given a cumulative distribution function F, define

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}$$

The reader may check that F^{-1} is an increasing, left-continuous function, and it is an inverse of F in the sense that

$$\{x; F^{-1}(x) \leq b\} = \{x; x \leq F(b)\}$$

F^{-1} is called the *quantile function*. If y is in the range of F, then $F(F^{-1}(y)) = y$.

We say that a random variable Z has the uniform distribution on $(0,1)$, or is uniformly distributed on $(0,1)$, if

$$\mathbb{P}(Z \in U) = \lambda(U \cap (0,1)),$$

where λ is Lebesgue measure. The canonical construction above is one way to construct uniform random variables.

Theorem 2.3.1 Let F be a *cumulative distribution function*. Let Z be uniformly distributed on $(0,1)$, and let $X := F^{-1}(Z)$. Then $F_X = F$.

Proof.

$$F_X(b) = \mathbb{P}\left(F^{-1}(Z) \leq b\right) = \lambda(Z \leq F(b)) = F(b).$$

□

Because of this theorem, in order to simulate a random variable with any given distribution, it suffices to have a good algorithm for simulating a uniformly distributed random variable.

These simple results have important consequences for how one defines random variable models in practice. It is rare to define a random variables or vectors by defining a specific function on an explicitly constructed probability space. Rather, a model will specify a cumulative distribution function or distribution measure. We know now there will be corresponding random variables. But the nature of probability underlying these random variables, and the way they are defined, is not restricted in any other way, and, in fact, is not usually relevant. The analyses we want to perform are always about probabilities of events defined in terms of the random variables, so they depend only on the distribution measure and not on the way the random variables are defined as functions on a probability space.

2.3.2 Kolmogorov's extension theorem

Recall that a *stochastic process* is a collection of random variables, $\{X_\alpha; \alpha \in \mathcal{A}\}$ on a common probability space, where $\mathcal{A}$ is some index set. The finite dimensional distribution of $\{X_\alpha; \alpha \in \mathcal{A}\}$, are the distribution measures.

$$\{\, \mathbb{F}_{X_{\alpha_1},...,X_{\alpha_m}};\ \alpha_1, ..., \alpha_m \in \mathcal{A},\ m \geq 1 \,\}$$

Just as with random vectors in the previous section, we want to define stochastic process models by specifying families of finite-dimensional distribution. This raises a generalization of the questions addressed in the previous section. Given an index set $\mathcal{A}$ and a family

$$\mathcal{J} = \{\, \mathbb{F}_{\alpha_1,...,\alpha_m};\ \alpha_1, ..., \alpha_m \in \mathcal{A},\ m \geq 1 \,\}$$

of distribution measures, when does there exist a stochastic process corresponding to this family in the sense that,

$$\mathbb{F}_{X_{\alpha_1},...,X_{\alpha_m}} = \mathbb{F}_{\alpha_1,...,\alpha_m} \tag{2.2}$$

for every finite subset $\{\alpha_1, ..., \alpha_m\} \in \mathcal{A}$.

To avoid any ambiguity, assume, by the *Axiom of Choice* if necessary, that $\mathcal{A}$ is ordered and that the family $\mathcal{J}$ is specified by $\mathbb{F}_{\alpha_1,...,\alpha_m}$ for $\alpha_1 < \alpha_2 < \cdots < \alpha_m$. If $S = \alpha_1, ..., \alpha_m$ is a finite, ordered subset of $\mathcal{A}$, let $\mathbb{F}_S$ denote $\mathbb{F}_{\alpha_1,...,\alpha_m}$. Similarly, let $x_S := (x_{\alpha_1}, ..., x_{\alpha_n})$ denote a point in $\mathbb{R}^S$.

Assume that a stochastic process $\{X_\alpha; \alpha \in \mathcal{A}\}$ satisfying (2.2) does exist. This forces $\mathcal{J}$ to satisfy a simple consistency property. Given any finite, ordered subset S and T of $\mathcal{A}$, with $S \subset T$, let $\Pi_{T|S}$ denote the projection of $\mathbb{R}^T$ onto $\mathbb{R}^S$:

$$\Pi_{T|S}(x_T) = x_S.$$

Then, since $\{X_\alpha\}_{\alpha \in S} := X_S = \Pi_{T|S}(X_T)$,

$$\mathbb{F}_S(U) = \mathbb{P}(X_S \in U) = \mathbb{P}\left(\Pi_{T|S}(X_T) \in U\right) = \mathbb{F}_T(\Pi^{-1}_{T|S}(U)).$$

This is the *Kolmogorov consistency criterion*. Wonderfully, it suffices for existence of $\{X_\alpha; \alpha \in A\}$.

Theorem 2.3.2 Let $\mathcal{J}$ be a consistent family of probability distribution functions, in the sense it satisfies *Kolmogorov consistency criterion*. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $\{X_\alpha; \alpha \in \mathcal{A}\}$ on it whose finite dimensional distributions are given by $\mathcal{J}$.

The proof of this theorem is not needed for the sequel. The strategy of the proof should be familiar by now. If we let $\Pi_{\mathcal{A}|T}$ denote the projection on $\mathcal{R}^{\mathcal{A}}$ onto $\mathbb{R}^T$, then $P\left(\Pi^{-1}_{\mathcal{A}|T}(U)\right) :=$

$\mathbb{F}_T(U)$, for finite $T \subset \mathcal{A}$ and $U \in \mathcal{B}(\mathbb{R}^T)$ defines a finitely additive probability measure on the cylinder sets of $\mathcal{R}^{\mathcal{A}}$. The consistency condition guarantees that this definition is consistent. Then *Carathéodory's extension theorem* is used to extend to a probability measure, call it $\mathbb{P}_{\mathcal{J}}$ on the σ-algebra of subsets of $\mathbb{R}^{\mathcal{A}}$ generated by cylinder sets. By this definition, if $T \subset \mathcal{A}$ is finite,

$$\mathbb{P}_{\mathcal{J}} \circ \Pi^{-1}_{\mathcal{A}|T} = \mathbb{F}_T.$$

Now, for $\omega = (\omega_\alpha)_{\alpha \in \mathcal{J}}$, let $X_\alpha(\omega) = \omega_\alpha$ pick out the α-coordinate of ω. Then the finite dimensional distribution of $\{X_\alpha\}$ are given by $\mathcal{J}$. Indeed, if $T = \{\alpha_1, ..., \alpha_m\}$,

$$\mathbb{F}_{X_{\alpha_1},...,X_{\alpha_m}}(U) = \mathbb{P}_{\mathcal{J}}\left(\Pi^{-1}_{\mathcal{A}|T}(U)\right) = \mathbb{F}_T(U).$$

The process X is called *canonical process* for the family $\mathcal{J}$.

2.4 Revisit Independence

As we discussed in the first chapter, the random variables $X_1, ..., X_n$ are *independent* if the corresponding σ-algebras $\sigma(X_1), ..., \sigma(X_n)$ are independent. Here independence is meant in the sense of *mutual independence*, **NOT** *pairwise independence*. The random variables of an arbitrary collection $\{X_\alpha; \alpha \in \mathcal{A}\}$ are *independent* if the random variables of any finite sub-collection are independent.

Example 2.4.1 Let $Y_1, Y_2, ...$ be the random variables defined in Example 2.1.5. We have already shown that each Y_i is discrete and $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = 0) = \frac{1}{2}$. We claim that $Y_1, Y_2, ...$ are independent. Notice that $\sigma(Y_i) = \{\{Y_i = 1\}, \{Y_i = 0\}, \Omega, \emptyset\}$ for each i. It suffices to show that $\{Y_1 = 1\}$ and $\{Y_2 = 1\}$ are independent. But $\mathbb{P}(Y_1 = 1, Y_2 = 1) = \lambda([\frac{3}{4}, 1]) = \frac{1}{4} = (\frac{1}{2})^2$, which proves independence.

Theorem 2.4.2 $X_1, ..., X_n$ are independent if and only if

$$\mathbb{P}(X_1 \in U_1, X_2 \in U_2, ..., X_n \in U_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in U_i)$$

for any Borel sets $U_1, ..., U_n$, if and only if

$$\mathbb{F}_{X_1,...,X_n} = \mathbb{F}_{X_1} \otimes \cdots \otimes \mathbb{F}_{X_n}.$$

Proof. Since $\{X_i \in U_i\} \in \sigma(X_i)$ for each i, independence of $X_1, ..., X_n$ implies the first equation by definition. Conversely, suppose the first equation holds, to prove $\{X_1 \in U_1\}, ..., \{X_n \in U_n\}$ are independent, we must show

$$\mathbb{P}\left(X_{i_1} \in U_{i_1}, ..., X_{i_r} \in U_{i_r}\right) = \prod_{k=1}^{r} \mathbb{P}(X_{i_k} \in U_{i_k})$$

for any $1 \leq_1 < \cdots < i_r \leq n$. But,

$$\mathbb{P}\left(X_{i_1} \in U_{i_1}, ..., X_{i_r} \in U_{i_r}\right) = \mathbb{P}\left(\{X_{i_k \in U_{i_k}}; 1 \leq k \leq r\} \cap \{X_j \in \Omega; j \notin \{i_1, ..., i_r\}\}\right)$$

The first equation implies that this equals

$$\prod_{k=1}^{r} \mathbb{P}(X_{i_k} \in U_{i_k}) \prod_{j \notin \{i_1,...,i_r\}} \mathbb{P}(X_j \in \Omega) = \prod_{k=1}^{r} \mathbb{P}(X_{i_k} \in U_{i_k}).$$

In terms of distribution measures,

$$\mathbb{F}_{X_1,...,X_n}(U_1 \times \cdots \times U_n) = \mathbb{F}_{X_1}(U_1) \times \cdots \times \mathbb{F}_{X_n}(U_n)$$

□

2.5 Expectation

Expected value is the name probabilists give to integration with respect to a probability measure. In the language of measure theory,

$$\textbf{??}\mathbb{E}[X] := \int_{\Omega} X(\omega)\mathbb{P}(d\omega) \tag{2.3}$$

We shall define the right hand side from scratch.

A random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said called *simple* if it has the form,

$$X(\omega) = \sum_{i=1}^{n} c_k \mathbf{1}_{A_k}(\omega),$$

where $A_1, ..., A_n \in \mathcal{F}$. A simple random variable is just a discrete random variable taking only a finite number of possible values. Conversely, if Y is a discrete random variable taking values in $\{ s_1, s_2, ..., s_N \}$,

$$Y(\omega) = \sum_{i=1}^{N} s_k \mathbf{1}_{\{Y=s_k\}}(\omega).$$

So it is *simple*. In this expression for Y, the event $\{Y = s_k\}$, $1 \leq k \leq N$ are disjoint. In an expression such as (**??**) for a simple random variable, A_1, ..., A_n are not necessarily disjoint.

The reason to introduce firstly the simple random variables is b that any random variable can be expressed as a point-wise limit of simple random variable. This undergoes two stages. The first step is achieved by the following easy lemma.

Lemma 2.5.1 Assume $X(\omega) \geq 0$ for all ω. There exists an increasing sequence, $\{X_n\}$, of *simple random variables* such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all ω.

Proof. Let's partition $[\,0,\,+\infty\,]$ as

$$[\,0,\,+\infty\,] = [\,0,\,n\,) \cup [\,n,\,+\infty\,] = \bigcup_{k=1}^{n2^n} [\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,) \cup [\,n,\,+\infty\,]$$

For any $\omega \in \Omega$ such that $X(\omega) \in [\,0,\,+\infty\,]$, $X(\omega) \in \bigcup_{k=1}^{n2^n} [\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,) \cup [\,n,\,+\infty\,]$. In other words, either $X(\omega) \in [\,n,\,+\infty\,]$ or $X(\omega) \in [\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,)$ for some k. Let's define *simple random variable*:

$$S_n(\omega) = \begin{cases} n, & \text{if } X(\omega) \in [\,n,\,+\infty\,]; \\ \frac{k-1}{2^n}, & \text{if } 0 \leq X(\omega) \in [\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,). \end{cases}$$

S_n can be rewritten as:

$$S_n(x) = n\mathbf{1}_{\{X^{-1}[\,n,\,+\infty\,]\}} + \frac{k-1}{2^n} \sum_{k=1}^{n2^n} \mathbf{1}_{\{f^{-1}[\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,)\}}$$

which is a non-negative, simple random variable. In fact, S_n is an increasing sequence of random variables, i.e., for fixed $\omega \in \Omega$, $S_{n+1}(\omega) \geq S_n(\omega)$, $\forall n$. We know that,

$$S_{n+1}(\omega) = \begin{cases} n+1, & \text{if } X(\omega) \in [\,n+1,\,+\infty\,]; \\ \frac{k-1}{2^{n+1}}, & \text{if } 0 \leq X(\omega) \in [\,\frac{k-1}{2^{n+1}},\,\frac{k}{2^{n+1}}\,). \end{cases}$$

For any fixed $\omega \in \Omega$ such that $X(\omega) \leq n+1$ (the case that $X(\omega) \geq n+1$ is trivial), it must happen that

$$X(\omega) \in [\,\frac{j-1}{2^n},\,\frac{j}{2^n}\,)$$

equivalently,

$$X(\omega) \in [\,\frac{2(j-1)}{2^n},\,\frac{2j}{2^n}\,)$$

Obviously, $X(\omega)$ can either locate within $[\,\frac{2(j-1)}{2^n},\,\frac{2j-1}{2^n}\,)$ or $[\,\frac{2(j-1)}{2^n},\,\frac{2j}{2^n}\,)$, under new partition. If $X(\omega) \in [\,\frac{2(j-1)}{2^n},\,\frac{2j-1}{2^n}\,)$, $S_{n+1}(\omega) = \frac{2(j-1)}{2^{n+1}} = S_n(\omega)$. And if $X(\omega) \in [\,\frac{2j-1}{2^n},\,\frac{2j}{2^n}\,)$, $S_{n+1}(\omega) = \frac{2j-1}{2^{n+1}} \geq \frac{2(j-1)}{2^{n+1}} = S_n(\omega)$. Indeed, $\{S_n\}$ is increasing.

Now, we are in the position to prove the convergence. Again, let's fix $\omega \in \Omega$, either $X(\omega) = +\infty$ or $0 \leq X(\omega) \leq n < n+1$. In the former case, $S_n(x) = n \to \infty$. In the latter case, $X(\omega) \in [\,\frac{k-1}{2^n},\,\frac{k}{2^n}\,)$, thus

$$|X(\omega) - S_n(\omega)| < \frac{1}{2^{n_0}} \text{ for some } n_0.$$

Thus, $\forall n \geq n_0$,

$$|X(\omega) - S_n(\omega)| < \frac{1}{2^n}$$

The assertion follows immediately. □

As the second step, let's decompose X to its positive part $X^+ = \max\{0, X\}$ and negative part $X^- = \min\{-X, 0\}$, i.e., $X = X^+ - X^-$. Then, if $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are not both infinity,

$$\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

The expectation is used to define other important quantities measuring the statistical properties of a random variable, such as, the n-th moment of a random variable, i.e, $\mathbb{E}[X^n]$. Let $\mu_X = \mathbb{E}[X]$, the n-th centered moment is $\mathbb{E}[(X - \mu_X)^n]$. The second centered moment is particularly important and is called the *variance*, we write

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

A simple calculation shows, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu_X^2$. Given two random variables with finite second moments, we also define the covariance:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Let's be aware of some simple properties of expectation.

Theorem 2.5.2 For X, Y random variables, $a, b \in \mathbb{R}$,

(i) *Linearity:* If $a\mathbb{E}[X] + \mathbb{E}[Y]$ makes sense (as an extended real number), $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$;
(ii) *Monotonicity:* If $Y \leq X$, $-\infty < \mathbb{E}[Y]$, then $\mathbb{E}[Y] \leq \mathbb{E}[X]$.

The idea of expectation is that $\mathbb{E}[X]$ is an *average value* of X. We see this most clearly for discrete random variables X. If X takes values in the set $\{ s_1, s_2, ...\}$, it follows from the definition that

$$\mathbb{E}[X] = \sum_k s_k \mathbb{P}(X = x_k).$$

Example 2.5.3 If X is $Ber(p)$, then $\mathbb{E}[X] = 0 \cdot (1 - p) + 1 \cdot p = p$ and $\mathrm{Var}(X) = p - p^2 = p(1 - p)$. If $X_1, X_2, ..., X_n$ are independent $Ber(p)$ random variables, then $Y = \sum_{i=1}^n X_i$ is $binomial(n, p)$ and $\mathbb{E}[Y] = \sum_{i=1}^n \mathbb{E}[X_i] = np$.

Example 2.5.4 If Z is a *Poisson* with parameter λ, then $\mathbb{E}[Z] = \lambda$. Here is a non-rigorous explanation. For large n, a $binomial(n, \lambda/n)$ random variables is approximately $Poisson(\lambda)$, by Theorem, with the approximation getting better as n gets larger. But for all n, the expectation of $binomial(n, \lambda/n)$ random variable is $n(\lambda/n) = \lambda$.

The important theorems for dealing with limits and expectations are those from *Real analysis*.

Theorem 2.5.5 *(Monotone Convergence Theorem)* Let $\{X_n\}$ be a sequence of random variables such that $0 \leq X_n \leq X_{n+1}$ for all n. Then,

$$\mathbb{E}[\lim_{n\to\infty} X_n] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Theorem 2.5.6 *(Fatou's Lemma)* Let $\{X_n\}$ be a sequence of almost surely positive random variables; that is $\mathbb{P}(X_n \geq 0) = 1$ for all n. Then

$$\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]$$

Theorem 2.5.7 *(Dominated Convergence Theorem)* Let $\{X_n\}$ be a sequence of random variables and assume that $\lim_{n\to\infty} X_n$ exists almost surely. Let X be a random variable such that $X = \lim_{n\to\infty} X_n$ almost surely. Assume that there is a random variable Z such that $|X_n| \leq Z$ for all n and $\mathbb{E}[Z] < +\infty$. Then

$$\mathbb{E}[X] = \mathbb{E}[\lim_{n\to\infty} X_n] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

We will use monotone convergence theorem to prove the most useful method for computing expectations of continuous random variables.

Theorem 2.5.8 Let $X_1, X_2, ..., X_n$ be random variables and let h be a *Borel* function such that $\mathbb{E}[h(X_1, ..., X_n)]$ is defined. Then

$$\mathbb{E}[h(X_1, ..., X_n)] = \int \cdots \int h(x_1, ..., x_n)\, d\mathbb{F}_{X_1,...,X_n}(x_1, ..., x_n). \tag{2.4}$$

In particular, if $X_1, ..., X_n$ has joint density f,

$$\mathbb{E}[h(X_1, ..., X_n)] = \int \cdots \int h(x_1, ..., x_n) f(x_1, ..., x_n)\, dx_1 \cdots dx_n. \tag{2.5}$$

Proof. (Outline) Firstly, we shall establish the theorem for simple functions h. This is just a matter of working through the definitions. For a general, positive h, let h_n be a sequence of simple functions increasing up to h. Then applying the monotone convergence on both sides of (2.4) for h_n implies the result for h. Finally, for general h, decompose it into its positive part and negative part. □

With this theorem we proceed to calculate the expected values of some continuous random variables.

Example 2.5.9 If X is uniform on $[\alpha, \beta]$, then

$$\mathbb{E}[X] = \frac{\alpha + \beta}{2}, \quad \text{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

If Y is exponential with parameter λ, then

$$\mathbb{E}[X] = \int_0^\infty x\lambda \exp -\lambda x dx = \frac{1}{\lambda}$$

If Z is normal with density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$, then

$$\mathbb{E}[Z] = \mu, \quad \text{Var}(X) = \sigma^2,$$

as the terminology of the normal density suggests.

2.6 Independence and Expectation

Expectations are particularly easy to work with when independence is around. Our first result is a generalization and development of Theorem 2.4.2.

Theorem 2.6.1 Random variables $X_1, ..., X_n$ are independent if and only if for any bounded Borel functions $h_1, ..., h_n$,

$$\mathbb{E}[h_1(X_1) \cdots h_n(X_n)] = \mathbb{E}[h_1(X_1)] \cdots \mathbb{E}[h_n(X_n)] \tag{2.6}$$

When $X_1, ..., X_n$ are independent, then this equation is true even when the h_i's are unbounded, as long as $\mathbb{E}[|h(X_i)|] < +\infty$ for all i.

Proof. If (2.6) is true, then it implies

$$\mathbb{P}\left(X_1 \in U_1, X_2 \in U_2, ..., X_n \in U_n\right) = \prod_{i=1}^n \mathbb{P}\left(X_i \in U_i\right)$$

by taking $h_i = \mathbf{1}_{U_i}$, for each i. Hence, by Theorem 2.4.2, $X_1, ..., X_n$ are independent. On the other hand, if $X_1, ..., X_n$ are independent, then we can derive (2.6) is an easy consequence of independence for simple random functions h_i. Then the general case is proved by the usual method of approximating general functions by simple functions. □

Remark 2.6.2 Random variables X and Y are said to be *uncorrelated* if $\text{Cov}(X, Y) = 0$.

Corollary 2.6.3 If X and Y are independent and have finite variance, then they are *uncorrelated*.

Proof. By previous theorem,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[X-\mu_X]\mathbb{E}[Y-\mu_Y] = 0.$$

□

When we are dealing with sums of random variables, we have,

Theorem 2.6.4 If $X_1, ..., X_n$ have finite variance,

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2 \sum_{1\le i\le j\le n} a_i a_j \mathrm{Cov}(X_i, X_j).$$

In particular, if $X_1, ..., X_n$ are *uncorrelated* or are *independent*, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i).$$

Example 2.6.5 Let $X_1, ..., X_n$ be independent $Ber(p)$ random variables. We know that $Y = \sum_{i=1}^{m} X_i$ is $binomial(n,p)$. Therefore,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p).$$

2.7 Moment Generating Functions; Introduction

To develop some applications of expectation we will introduce moment generating functions in this section.

Definition 2.7.1 If X is a random variable, the function,

$$\Phi_X(\lambda) = \mathbb{E}[\exp i\lambda X],\ \lambda \in \mathbb{R},$$

is called the *characteristic function* of X. The function,

$$m_X(t) = \mathbb{E}[e^{tX}],\ t \in \mathbb{R},$$

is called the *moment generating function* of X.

In terms of distribution measure,

$$\Phi_X(\lambda) = \int \exp i\lambda x d\mathbb{F}_X(x), \text{ and } m_X(t) = \int e^{tx} d\mathbb{F}_X(x).$$

In particular, $\Phi_X(\lambda)$ is the *Fourier transform* of the law $\mathbb{F}_X$ of X. We shall discuss the characteristic function in depth later on. For now, we will be more interested in moment generating function. The following properties of moment generating function are important.

Theorem 2.7.1 Assume that there is an $\epsilon > 0$ such that $m_X(t) < +\infty$ for all t satisfying $-\epsilon < t < \epsilon$. Then $\mathbb{E}[|X|^n]$ is finite for all n and

$$\mathbb{E}[|X|^b] = \frac{d^n m_X(t)}{dt^n}|_{t=0}$$

Theorem 2.7.2 Let $X_1, ..., X_n$ be independent, and let $S = X_1 + \cdots + X_n$. Then the moment generating function $m_S(t)$ is the product of the moment generating functions $m_{X_i}(t)$ of the summands:

$$m_S(t) = m_{X_1}(t) m_{X_2}(t) \cdots m_{X_n}(t)$$

Likewise, the characteristic function $\Phi_S(\lambda)$ of S is the product of characteristic functions:

$$\Phi_S(t) = \Phi_{X_1}(t) \Phi_{X_2}(t) \cdots \Phi_{X_n}(t).$$

Example 2.7.3 Direct calculation shows that if X is a $Ber(p)$ random variable, $m_X(t) = 1 - p + pe^t$. Therefore, if $X_1, ..., X_n$ are independent $Ber(p)$, then the moment generating function of the $binomial(n, p)$ random variable is

$$m_Y(t) = [1 - p + pe^t]^n.$$

Remark 2.7.4 We will see later that, as long as the moment generating function is defined on a non-trivial interval of t-values, it uniquely determines the distribution of the corresponding random variable. This is enormously useful for studying sums of independent random variables. Even if we cannot derive an exact formula for the distribution of sum of independent random variables, we are often able to obtain a lot of information from the moment generating function of the sum, which we can compute.

2.8 Convergence of Random Variables

There are several senses in which a sequence of random variables can converge to another random variable. All are important. We have already encountered *almost sure convergence*. We say that $\{X_n\}$ converges to X almost surely, which is written sometimes as $X_n \overset{\text{a.s.}}{\to} X$ and sometimes $\lim X_n \overset{\text{a.s.}}{=} X$ if the $\lim X_n(\omega)$ exists on a set of probability one and, on the set where it exists, equals X with probability one.

On the other hand, $\{X_n\}$ *converges in probability* to X, written sometimes as $X_n \overset{P}{\to} X$, if for every $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n - X| > \epsilon\right) = 0.$$

This definition is framed for limit X which can not $+\infty$ or $-\infty$. In measure theory, convergence in probability is called *convergence in measure*.

Let $p > 0$. The sequence $\{X_n\}$ *converges in* L^p, written $X_n \xrightarrow{L^p} X$ if

$$\lim_{n\to\infty} \mathbb{E}\big[|X_n - X|^p\big] = 0.$$

A final sense of convergence, but one we will put off studying for a while, is *convergence in distribution*, sometimes called *convergence in law*. X_n converges in distribution to X as $n \to \infty$ if for every bounded continuous function f,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$$

This definition will appear unmotivated to the non-initiated. We will explain in a later chapter what it means. For the moment, notice that convergence in distribution is quite different in one respect from the other types of convergence. Almost-sure, probability and L_p convergence all require that the random variables be defined on one probability space. This is not true of convergence in distribution. It is purely a statement about distribution measure, because it can be written as

$$\lim_{n\to\infty} \int f(x) d\mathbb{F}_{X_n}(x) = \int f(x) d\mathbb{F}_X(x)$$

for all bounded, continuous f. This is a form of what analysts would call *weak convergence*, or more accurately *weak*-convergence*. One can view a probability distribution measure $\mathbb{F}$ as a linear operator $f \mapsto \int f(x)\, d\mathbb{F}(x)$, on bounded, continuous functions on $\mathbb{R}$, and convergence in distribution is a type of convergence of these operators.

To discuss the relation between convergence in L^p and convergence in probability, we introduce one of the most basic and simple tools of probability.

Theorem 2.8.1 (i) *Markov's Inequality:* If $X \geq 0$, almost surely, then for all $a > 0$,

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}[X\mathbf{1}_{\{X\geq a\}}]}{a} \leq \frac{\mathbb{E}[X]}{a}.$$

(ii) *Chebyshev's inequality:* If $a > 0$, then

$$\mathbb{P}\left(|X - \mu_X| \geq a\right) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

(iii) *Generlization:* If $a > 0$,

$$\mathbb{P}\left(|X - Y| \geq a\right) \leq \frac{\mathbb{E}[|X - Y|^p]}{a^p}$$

(iv) *Exponential bounds:* For $a > 0$,

$$\mathbb{P}(X > a) \leq \exp\left(-\sup_{t>0}(ta - \log M_X(t))\right)$$

Proof. For (i), simply use the fact that $\mathbb{P}(X > a) = \mathbb{E}[\mathbf{1}_{\{X \geq a\}}]$ and that

$$\mathbf{1}_{\{X \geq a\}} \leq \frac{X}{a} \mathbf{1}_{\{X \geq a\}}.$$

Moreover,

$$a\mathbb{P}(|X| \geq a) \leq \mathbb{E}[|X| \mathbf{1}_{\{|X| \geq a\}}]$$

By *dominated convergence theorem*,

$$\lim_{a \to \infty} a\mathbb{P}(|X| \geq a) = 0$$

Thus, *Markov's inequality* also tells us the rate of convergence as $a \to \infty$, i.e.,

$$\mathbb{P}(|X| \geq a) = o(\frac{1}{a})$$

where $o(y)$ stands for the higher order terms w.r.t. y. For (ii), note that

$$\mathbb{P}\left(|X - \mu_X| \geq a\right) = \mathbb{P}\left(|X - \mu_X|^2 \geq a^2\right),$$

and apply *Markov's inequality*. Part (iii) is also proved by a simple application of Markov inequality.

For part (iv), consider

$$\mathbb{P}(X > a) = \mathbb{P}(\exp(tX) > \exp(ta))$$

for any $t > 0$, apply *Markov's inequality*,

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[\exp(ta)]}{\exp ta} = \exp(-ta + \log m_X(t)).$$

To optimize the right hand side,

$$\mathbb{P}(X > a) \leq \exp\{-(ta - \log m_X(t))\} \leq \exp\{-\sup_{t>0}(ta - \log m_X(t))\}$$

□

The application of exponential bounds to i.i.d. sequence leads to *Chernoff's bounds*. We are not interested in great generality but a special application of such bounds.

Corollary 2.8.2 Let $X_1, ..., X_n$ be independent *Bernoulli* random variables, $\mathbb{P}(X_i = 0) = 1 - p$ $\mathbb{P}(X_i = 1) = p$, then $\bar{X}_n \to p$ in probability.

Proof. Let's look at:

$$\log m_X(t) = \log[(1-p) + pe^t]$$

Notice

$$\sup_{t>0}[at - \log m_{\bar{X}_n}(t)] = a \ln \frac{a}{p} + (1-a) \ln \frac{1-a}{1-p} \geq 2(a-p)^2$$

Let $a > p$, by the exponential bounds,

$$\mathbb{P}(\bar{X}_n > a + \epsilon) \leq \exp(-2n\epsilon^2)$$

By symmetry, if we exchange the role, i.e., $Y_i = 1 - X_i$, then

$$\mathbb{P}(\bar{X}_n < p - \epsilon) \leq \exp(-2n\epsilon^2)$$

Thus,

$$\mathbb{P}(|\bar{X}_m| > \epsilon) \leq 2\exp(-2n\epsilon^2).$$

As $n \to \infty$, we obtain $\bar{X}_n \to p$ in probability. □

Theorem 2.8.3 (i) If $X_n \overset{\text{a.s.}}{\to} X$, then $X_n \overset{P}{\to} X$;
(ii) If $X_n \overset{L^p}{\to} X$, then $X_n \overset{P}{\to} X$.

Proof. Part (ii) is an immediate consequence of *generalization inequality* and the definition of convergence in probability. To prove part (i), write

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{E}[\mathbf{1}_{\{|X_n - X| \geq \epsilon\}}].$$

Notice that if $X_n \overset{\text{a.s.}}{\to} X$, then $\mathbf{1}_{\{|X_n - X| \geq \epsilon\}} \overset{\text{a.s.}}{\to} 0$, whenever $\epsilon > 0$. Now apply the dominated convergence theorem. □

Remark 2.8.4 As a final remark for this chapter, we will see later that convergence in probability implies convergence in distribution, and it follows then that almost sure and L^p convergence also imply convergence in distribution.

Chapter 3

Large Number of Sequence of Random Variables

Let $X_1, X_2, ...$ be a sequence of random variables, and, for every n, define the empirical average of $X_1, ..., X_n$ as

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{i} X_i.$$

In this chapter, we shall study large number laws for the sequence of empirical averages. We say that a *weak law of large number* hods for the empirical average if there is a deterministic sequence $\{b_n\}$ such that

$$\bar{X}_n - b_n \to 0 \text{ in probability as } n \to \infty.$$

We say that a *strong law of large numbers* holds if

$$\bar{X}_n - b_n \to 0 \text{ almost surely as } n \to \infty.$$

In the classical statements of large number laws, the random variables have an identical and finite mean $\mu = \mathbb{E}[X_i]$, and $b_n = \mu$ for every n. This is the case, in particular, when the random variables have identical laws.

In this chapter, we state and prove some basic strong and weak laws. Our interest lies as much with the techniques of proof as with the results. We shall see that there are only two simple tools, *Chebyshev's inequality* and the *Borel-Cantelli lemma*, but by skillfully combining them, we can derive some deep results. We will also study the different but related question of convergence of a random series $\sum_{i=1}^{\infty} Z_i$ for a sequence of independent random variables $\{Z_n\}$. We state *Kolmogorov's three series theorem* giving necessary and sufficient conditions for convergence and prove sufficiency but not necessity. A second proof of the strong law of large number is then presented.

Throughout the chapter we use the standard abbreviation *i.i.d.* to denote *independent and identically distributed*. A limit in probability of X_n is often denoted $(P)\lim_{n\to\infty} X_n$.

3.1 The Weak Law of Large Numbers

Theorem 3.1.1 Let $\{X_n\}$ be a sequence of uncorrelated random variables with common mean μ and variance σ_n^2 such that $\max_n \sigma_n^2 = K < \infty$; (For every i, $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.) Then

$$(P)\ \lim_{n\to\infty} \bar{X}_n = \mu. \tag{3.1}$$

Proof. To simplify notation we shall assume that $\mu = 0$. This entails no loss of generality since (3.1) is equivalent to $(P)\ \lim(1/n)\sum_{i=1}^n (X_i - \mu) = 0$ and $X_i - \mu$ has mean 0. From *Chebyshev's inequality*, we obtain:

$$\mathbb{P}\left(|\bar{X}_n| > \epsilon\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} \leq \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{n^2\epsilon^2} \leq \frac{K}{n\epsilon^2}. \tag{3.2}$$

Thus, for every $\epsilon > 0$, $\mathbb{P}\left(|\bar{X}_n| > \epsilon\right) \to 0$ as $n \to \infty$, proving convergence in probability. □

Remark 3.1.2 We have actually showed a stronger result, namely that $\bar{X}_n$ convergence in L^2 to μ.

The argument we used in the proof extends easily to more general situation. For example, let $X_1, X_2, ...$ be uncorrelated with respective means $\mu_1, \mu_2, ...$, and assume that $\sup_n \mathrm{Var}(X_n) < +\infty$. Then,

$$(P)\ \lim_{n\to\infty} \bar{X}_n - \frac{1}{n}\sum_{i=1}^n \mu_i = 0$$

We emphasize the bound (3.3) based on *Chebyshev's inequality* for the probability deviation of the empirical mean from the true mean. for the case of uncorrelated random variables with arbitrary means and common variance σ^2, we restate it as:

$$\mathbb{P}\left(|\bar{X}_n - \frac{1}{n}\sum_{i=1}^n \mu_i| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{3.3}$$

This bound is ultimately the basis for all our proofs of large number laws.

3.2 The Borel-Cantelli Lemma

To pass from weak laws to strong laws of large numbers, we need a tool for passing from convergence in probability to almost sure convergence. The *Borel-Cantelli lemma* serves precisely this function. Actually, the *Borel-Cantelli lemma* has two parts, the second of which, concerning independent events, is not directly relevant to our treatment of large number laws. Nevertheless, we include its statement and proof.

Let $\{A_n\}$ be a countable sequence of events, and define event

$$\begin{aligned}&\{\{A_n\} \text{ occurs infinitely often}\}\\ :=&\{\,\omega \,|\, \exists \text{ a subsequence } \{n_k\} \text{ (depending on } \omega) \text{ such that } \omega \in A_{n_k},\, \forall k\,\}.\end{aligned}$$

We shall use $\{\{A_n\}\, i.o.\,\}$ for $\{\{A_n\} \text{ occurs infinitely often}\}$. An equivalent definition is:

$$\{\{A_n\}\, i.o.\,\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$$

In the set theory $\{\{A_n\}\, i.o.\,\}$ is denoted $\limsup A_n$. The complementary set construction to the lim sup is

$$\liminf A_n := \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n.$$

Notice that

$$\{\ A_n \text{ occurs finitely often}\} := \{\{A_n\}\, i.o.\,\}^{\mathcal{C}} = \liminf A_n^{\mathcal{C}}.$$

Example 3.2.1 This example shows the immediate relevance of the $\limsup A_n$ construction to questions of convergence. Let $X_1, X_2, ...$ and Y be random variables. By the statement $\lim X_n(\omega) \neq Y(\omega)$, we mean either $\lim X_n(\omega)$ does not exist, or the limit exists and is not equal to $Y(\omega)$. Now suppose that $\lim X_n(\omega) \neq Y(\omega)$. Then given any $\epsilon > 0$ there is a subsequence $n_1 < n_2 < \cdots$ along which $|X_{n_k}(\omega) - Y(\omega)| > \epsilon$, for every k. Conversely, if there exists such an ϵ and such a subsequence, then $\lim X_n(\omega) \neq Y(\omega)$. Thus,

$$\{\,\omega \,|\, \lim X_n(\omega) \neq Y(\omega)\,\} = \bigcup_{j=1}^{\infty} \{|X_n - Y| > \frac{1}{j}\, i.o.\,\}. \tag{3.4}$$

Similarly, let $\epsilon_n \downarrow 0$ be a sequence of positive integers decreasing down to 0. Then

$$\{\lim X_n \neq Y\} \subset \{|X_n - Y| > \epsilon_n \text{1}.o.\,\} \tag{3.5}$$

The *Borel-Cantelli lemma* gives simple conditions for concluding when the probability of $\{\,\{A_n\}\, i.o.\,\}$ is 0 or 1.

Proposition 3.2.2 *(Borel-Cantelli Lemma)* Let $\{A_n\}$ be a sequence of events,

(i) If $\sum_{i=1}^{n} \mathbb{P}(A_n) < +\infty$, then

$$\mathbb{P}(\{\,\{A_n\}\, i.o.\,\}) = 0.$$

(ii) If the events $A_1, A_2, ...$ are independent and $\sum_{i=1}^{\infty} \mathbb{P}(A_n) = \infty$, then

$$\mathbb{P}(\{\,\{A_n\}\, i.o.\,\}) = 1.$$

Proof. The proof for (i) uses the sub-additivity of $\mathbb{P}$. For every m,

$$\mathbb{P}(\{\,\{A_n\}\, i.o.\}) = \mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\right) \leq \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n\right) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n).$$

The condition $\sum_{i=1}^{\infty} \mathbb{P}(A_n) < +\infty$ thus implies that

$$\mathbb{P}(\{\,\{A_n\}\, i.o.\}) \leq \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(A_n) = 0.$$

The proof of (ii) is more involved. We shall do it by showing that $\mathbb{P}(\{\,\{A_n\}\, i.o.\}^{\mathcal{C}}) = 0$. First notice that because of the continuity of the probability under limits of increasing and decreasing sets

$$\mathbb{P}(\{\,\{A_n\}\, i.o.\,\}^{\mathcal{C}}) = \mathbb{P}\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^{\mathcal{C}}\right) = \lim_{m\to\infty} \lim_{N\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{N} A_n^{\mathcal{C}}\right) \tag{3.6}$$

Now we use the independence of A_1, A_2, ... and the inequality $\log(1-x) < -x$, $0 < x < 1$, to obtain

$$\begin{aligned}\log \mathbb{P}\left(\bigcap_{n=m}^{N} A_n^{\mathcal{C}}\right) &= \log \prod_{n=m}^{\infty} \mathbb{P}(A_n^{\mathcal{C}}) = \sum_{n=m}^{N} \log \mathbb{P}(A_n^{\mathcal{C}}) \\ &= \sum_{n=m}^{N} \log(1 - \mathbb{P}(A_n)) \leq -\sum_{n=m}^{N} \mathbb{P}(A_n).\end{aligned}$$

The assumption that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ therefore implies that

$$\lim_{N\to\infty} \log \mathbb{P}(\bigcup_{n=m}^{N} A_n^{\mathcal{C}}) \leq -\lim_{N\to\infty} \sum_{n=m}^{N} \mathbb{P}(A_n) = -\infty,$$

and thus $\lim_{N\to\infty} \mathbb{P}(\bigcup_{n=m}^{N} A_n^{\mathcal{C}}) = 0$ for every $m \geq 1$. Statement (ii) follows from this conclusion and equation (3.6). □

We give a preliminary application of the *Borel-Cantelli lemma*, part (i), to a result on the relationship between almost sure convergence and convergence in probability.

Proposition 3.2.3 (i) If X_n converges to Y in probability the there exists a subsequence $\{X_{n_k}\}$ such that $(a.s.)\lim X_{n_k} = Y$;
(ii) Suppose that $\{X_n\}$ is a sequence of random variables such that every subsequence $\{X_{n_k}\}$ contains a sub-subsequence converging almost surely to the random variable Y. Then $(P)\lim X_n = Y$.

Proof. It suffices to consider the case in which $Y = 0$. Thus, suppose that $(P)\lim X_n = 0$. Choose a subsequence $\{n_k\}$ of the integers such that for every k,

$$\mathbb{P}(|X_{n_k} > \frac{1}{k^2}|) < \frac{1}{k^2}.$$

Since $\sum_k k^{-2} < +\infty$, the *Borel-Cantelli lemma* implies that

$$\mathbb{P}(|X_{n_k}| > \frac{1}{k^2}\, i.o.\,) = 0.$$

By (3.5),

$$\mathbb{P}(\lim_{k\to\infty} X_{n_k} \neq 0) = 0,$$

Thus proving (i).

A proof by contradiction, using the fact that almost sure convergence implies convergence in probability, proves part (ii). □

3.3 Strong Large Number Laws

The technique used here to obtain strong law rests on the following observation.

Lemma 3.3.1 Let $\{Y_k\}$ be sequence of random variables, and suppose that for every $\epsilon > 0$, $\mathbb{P}(|Y_k| > \epsilon\ i.o.\,) = 0$. Then $(a.s.)\lim Y_k = 0$.

Proof. By using identity of (3.4) and the hypothesis of the lemma,

$$\begin{aligned}\mathbb{P}(\lim_{k\to\infty} Y_k \neq 0) &= \mathbb{P}\left(\bigcup_{k=1}^{\infty}\{|Y_k| > \frac{1}{j}\ i.o.\,\}\right)\\ &\leq \sum_{j=1}^{\infty}\mathbb{P}(|Y_k| > \frac{1}{j}\ i.o.\,) = 0.\end{aligned}$$

□

The use of the Lemma above is simple. To show $(a.s.)\lim Y_k = 0$, we first employ *Chebyshev's inequality* to bound $\mathbb{P}(|Y_k| > \epsilon)$. If for every $\epsilon > 0$, $\sum_{k=1}^{\infty} \mathbb{P}(|Y_k| > \epsilon) < \infty$, then the *Borel-Cantelli lemma* implies that for every $\epsilon > 0$,

$$\mathbb{P}(|Y_k| > \epsilon \ i.o.\ ) = 0.$$

By this lemma, the almost sure convergence of $\{Y_n\}$ to 0 as $n \to \infty$ follows. Here are two simple applications of this strategy. The first is a proof of the strong law of large numbers for *Bernoulli* random variables, stated previously in Chapter 1.

Corollary 3.3.2 Let X_1, X_2,... be i.i.d. *Bernoulli* random variables with distribution $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$. Then $(a.s.)\lim \bar{X}_n = p$.

Proof. From the corollary of *Chernoff's inequality* in *Chapter 2*,

$$\sum_{n=1}^{\infty} \mathbb{P}(|\bar{X}_n - p| > \epsilon) \leq \sum_{n=1}^{\infty} 2\exp\left(-2n\epsilon^2\right) < \infty,$$

for every ϵ. The strong law follows immediately from Lemma 3.3.1 and the *Borel-Cantelli* lemma. □

Corollary 3.3.3 Let $X_1, X_2, ...$ be uncorrelated random variables with identical mean μ and identical variance $\sigma^2 < +\infty$. Then $(a.s.)\lim_{k\to\infty} \bar{X}_{k^2} = \mu$.

Proof. By the inequality (3.3),

$$\mathbb{P}\left(|\bar{X}_{k^2} - p| > \epsilon\right) \leq \frac{\sigma^2}{k^2\epsilon^2}$$

Since $\{k^{-2}\}_{k\geq 1}$ is a summable sequence, the corollary follows from Lemma 3.3.1 and *Borel-Cantelli*. □

We can go much beyond these naive application of *Borel-Cantelli*. The first result we prove is the strong law for uncorrelated random variables. The idea is to first extract subsequence of $\{\bar{X}\}$ that converges almost surely. This we have already done in above corollary. Then we show $\bar{X}_n$ at values of n between successive points of the subsequence, again using the *Borel-Cantelli* machine.

Theorem 3.3.4 Let $\{X_n\}$ be a sequence of uncorrelated random variables with common mean μ and common variance $\sigma^2 < \infty$. Then $(a.s.)\lim \bar{X} = \mu$.

Proof. Without loss of generality we may assume that $\mu = 0$l otherwise we just replace $X_i$by $X_i - \mu$. We know already from above corollary that

$$(a.s.)\lim \bar{X}_{k^2} = 0. \tag{3.7}$$

Now, consider n and k such that $k^2 < n < (k+1)^2$. We wish to compare $\bar{X}_n$ to $\bar{X}_k^2$. For this purpose, define

$$W_k = \frac{1}{k^2} \max\{|\sum_{k^2+1}^{n} X_i| : k^2 < n < (k+1)^2\}$$

For n in this range, we have

$$\begin{aligned}\bar{X}_n &= \frac{1}{n}\sum_{i=1}^{k^2} X_i + \frac{1}{n}\sum_{i=k^2+1}^{n} X_i \\ &= \frac{k^2}{n}\bar{X}_{k^2} + \frac{k^2}{n}(\frac{1}{k^2}\sum_{i=k^2+1}^{n} X_i),\end{aligned}$$

from which it follows that

$$|\bar{X}_n| \leq |\bar{X}_{k^2}| + W_k \tag{3.8}$$

If we can show that $(a.s.) \lim W_k = 0$, the strong law of large numbers will clearly follow from (3.7) and (3.8).

To show that, it suffices to prove $\mathbb{P}(W_k > \epsilon \; i.o.\,) = 0$ for any $\epsilon > 0$ because of Lemma 3.3.1. By definition of W_k,

$$\{W_k > \epsilon\} = \bigcup_{n=k^2+1}^{(k+1)^2-1} \{|\sum_{i=k^2+1}^{n} X_i| > k^2\epsilon\}. \tag{3.9}$$

Notice that for $k^2 < n < (k+1)^2$,

$$\mathrm{Var}\left(\sum_{i=k^2+1}^{n} X_i\right) = (n-k^2)\sigma^2 \leq 2k\sigma^2. \tag{3.10}$$

From (3.9), (3.10) and *Chebyshev's inequality*,

$$\mathbb{P}(W_k > \epsilon) \leq \sum_{i=k^2+1}^{(k+1)^2-1} \frac{2k\sigma^2}{k^4\epsilon^2} \leq \frac{(2k)^2\sigma^2}{k^4\epsilon^2} = \frac{4\sigma^2}{k^2\epsilon^2},$$

where the second inequality follows from the fact that the sum contains $2k$ terms. Since $\sum k^{-2} < +\infty$, the *Borel-Cantelli lemma* implies that $\mathbb{P}(W_k > \epsilon \, i.o.\,) = 0$, as desired. □

The next step in pushing forward the strong law is to relax the assumption that second moments being bounded. This was first achieved by *Kolmogorov*, who dropped the second moment condition at the expense of assuming that the random variables being independent and identically distributed.

Theorem 3.3.5 *(Strong Law of Large Numbers)* Let $\{X_n\}$ be a sequence of independent, identically distributed, integrable random variables with mean μ. Then

$$(a.s.) \lim_{n\to\infty} \bar{X}_n = \mu.$$

We will not present *Kolomogorov's* proof here. Instead, we will prove the following strong theorem due to *Etemadi*. We follow the treatment in *Probability and Measre*, *P.Billingsley*. The rest of this section will be devoted to the proof of this theorem. Again, the proof rests ultimately on *Chebyshev's inequality* and *Borel-Cantelli lemma*. However, we need another, important technique – truncation – and the concept of equivalent sequences.

Definition 3.3.1 Two sequences of random variables $\{U_n\}$ and $\{V_n\}$ are equivalent, written $\{U_n\} \sim \{V_n\}$, if $\sum_{n=1}^{\infty} \mathbb{P}(U_n \neq V_n) < \infty$.

It is immediate from the *Borel-Cantelli lemma* that if $\{U_n\} \sim \{V_n\}$,

$$\mathbb{P}(U_n \neq V_n \ i.o.\,) = 0, \text{ or, equivalently, } \mathbb{P}(\{U_n \neq V_n \ i.o.\,\}^{\mathcal{C}}) = 1.$$

In other words, because

$$\{U_n \neq V_n \ i.o.\,\}^{\mathcal{C}} = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{U_n = V_n\}, \tag{3.11}$$

equivalent sequences $\{U_n\}$ and $\{V_n\}$ eventually coincide with probability 1. Thus, in discussing tail events we may replace $\{U_n\}$ by $\{V_n\}$ and vice-versa. For example, we have

Lemma 3.3.6 Suppose $\{U_n\} \sim \{V_n\}$. Then $\{\lim \bar{U}_n \text{ exists }\} = \{\lim \bar{V}_n \text{ exists}\}$ and $\lim \bar{U}_n = \lim \bar{V}_n$, $\mathbb{P}$-almost surely on the set $\{\lim \bar{U}_n \text{ exists }\}$. In particular, if Z is a random variable, then $(a.s.) \lim \bar{U}_n = Z$ if and only if $(a.s.) \lim \bar{V}_n = Z$. Also, $(P) \lim \bar{U}_n = Z$ if and only if $(P) \lim \bar{V}_n = Z$.

Remark 3.3.7 The claim about convergence in probability follows from the claim that almost sure limits of equivalent sequences are almost surely equal and the characterization of convergence in probability in part (ii) of Proposition 3.2.3.

Now let $\{X_n\}$ be a given sequence of pairwise independent, identically distributed random variables with finite mean μ. The first step in the proof of the strong law is to replace $\{X_n\}$ by an equivalent sequence $\{Y_n\}$ such that each Y_n is bounded. Then we can apply *Chebyshev's inequality* to these bounded random variables and obtain a weak law of large numbers for the equivalent sequence. Finally, we pass from the weak law to strong law for the equivalent sequence by using *Borel-Cantelli lemma* along appropriate subsequence. The strong law for the original sequence $\{X_n\}$ then follows from the lemma right above.

To construct an equivalent sequence, we truncate each X_n at level n; that is, we define $Y_n = X_n \mathbf{1}_{\{|X_n| \leq n\}}$ for each positive integer n. This truncation produces an equivalent

sequence precisely because we are assuming that X_i's have finite mean. Indeed, from the identity, $\mathbb{E}[|X|] = \int_0^\infty \mathbb{P}(|X| > x)dx$, which is obtained by integration by parts, we get that

$$\mathbb{E}[|X|] < +\infty \text{ if and only if } \sum_{n=1}^{\infty} \mathbb{P}(|X| < n) < +\infty$$

Thus, since the X_n, $n \geq 1$, are identically distributed,

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{i=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty,$$

and it follows that $\{Y_n\} \sim \{X_n\}$. Note that the independence assumption is not used in this argument. Now, let $\mu_n = \mathbb{E}[Y_n]$ and define $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$. Notice that

$$\lim_{n\to\infty} \mu_n = \mu \text{ and hence } \lim_{n\to\infty} \bar{\mu}_n = \mu.$$

We shall prove a weak law for $\{Y_n\}$ using the assumption of pairwise independence of $X_1, X_2, ...$ In what follows $\bar{Y}_n$ is the empirical mean of Y_1,..., Y_n.

Lemma 3.3.8 If $\{X_n\}$ is a sequence of pairwise independent identically distributed random variables with finite mean μ, then

$$\mathbb{P}(|\bar{Y}_n - \bar{\mu}_n| > \epsilon) \leq \frac{1}{\epsilon^2 n}\mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \leq n\}}], \tag{3.12}$$

and hence $(P)\lim \bar{Y}_n = \mu$.

Proof. Since $X_1, X_2, ...$ are pairwise independent, $Y_1, Y_2, ...$ are uncorrelated. Thus,

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_n) &= \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}(Y_k) \\
&\leq \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}[X_k^2 \mathbf{1}_{\{|X_k| \leq k\}}] \\
&\leq \frac{1}{n^2}(n\mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \leq n\}}]) = \frac{1}{n}\mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \leq n\}}].
\end{aligned}$$

An application of dominated convergence theorem shows that the right hand side converges to 0 as $n \to \infty$, and hence that $(P)\lim \bar{Y}_n - \bar{\mu}_n = 0$. The assertion follows immediately. □

We turn now to the completion of the proof of *Etemadi's* strong law. First note that it suffices to consider only the case in which the random variables are almost surely non-negative, i.e., $\mathbb{P}(X_n \geq 0) = 1$. To see why, write $X_n = X_n^+ = X_n^-$. The sequence $\{X_n^+\}$ and $\{X_n^-\}$ both satisfy the hypothesis of Lemma 3.3.8 and consist of non-negative random variables. Since $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^+ - \frac{1}{n}\sum_{i=1}^{n} X_i^-$, it suffices to establish Lemma 3.3.8 for them

separately in order to prove it for $\{X_n\}$. The assumption of positivity will be technically convenient, and we shall henceforth assume it is satisfied.

Since $\{Y_n\} \sim X_n$, we know from Lemma 3.3.6 that to complete the proof of the strong law, it suffices to prove $(a.s.) \lim \bar{Y}_n = \mu$. The proof proceeds in two steps. Step 1 is the next lemma establishing almost sure convergence of a geometrically increasing subsequence. For positive numbers a, let $\lfloor a \rfloor$ denote the greatest integer less than or equal to a.

Lemma 3.3.9 For every $a > 1$, we have that $(a.s.) \lim \bar{Y}_{\lfloor a^n \rfloor} = \mu$.

The second step is to deduce the strong law from Lemma 3.3.9. Here is where we use the assumed non-negativity of the X_n.

Lemma 3.3.10 If the random variables X_n's are non-negative, the statement $(a.s.) \lim \bar{Y}_n = \mu$ for every $a > 1$ implies $(a.s.) \lim \bar{Y}_n = \mu$.

We prove Lemma 3.3.10 first, as it is easier. Thus, assume $(a.s.) \lim \bar{Y}_{\lfloor a^n \rfloor} = \mu$.

Proof. Let $\lfloor a^n \rfloor < k < \lfloor a^{n+1} \rfloor$. Since the Y_k random variables are non-negative,

$$\frac{\lfloor a^n \rfloor}{\lfloor a^{n+1} \rfloor} \bar{Y}_{\lfloor a^n \rfloor} = \frac{1}{\lfloor a^{n+1} \rfloor} \sum_{i=1}^{\lfloor a^n \rfloor} Y_i < \frac{1}{k} \sum_{i=1}^{k} Y_i = \bar{Y}_k \leq \frac{1}{\lfloor a^n \rfloor} \sum_{i=1}^{\lfloor a^{n+1} \rfloor} Y_i = \frac{\lfloor a^{n+1} \rfloor}{\lfloor a^n \rfloor} \bar{Y}_{\lfloor a^{n+1} \rfloor}.$$

By letting $n \to \infty$ in this string of equalities and inequalities and using $\lim \lfloor a^{n+1} \rfloor / \lfloor a^n \rfloor = a$, it follows that for every $a > 1$,

$$\mathbb{P}(\frac{\mu}{a} \leq \liminf \bar{Y}_k \leq \limsup \bar{Y}_k \leq a\mu) = 1.$$

Letting $a \downarrow 1$ gives $\mathbb{P}(\lim \bar{Y}_k = \mu) = 1$, as desired. □

We are only left to prove Lemma 3.3.9:

Proof. We shall need the following simple, analytic fact. For each $a > 1$, there is a constant C_a such that

$$\sum_{n=1}^{\infty} \frac{1}{\lfloor a^n \rfloor} \mathbf{1}_{\{x \leq \lfloor a^n \rfloor\}} \leq \frac{C_a}{x} \text{ for all } x > 0. \tag{3.13}$$

Indeed, let $N_x = \min\{\, n \,|\, \lfloor a^n \rfloor \geq x \,\}$. Then

$$\sum_{n=1}^{\infty} \frac{1}{\lfloor a^n \rfloor} \mathbf{1}_{\{x \leq \lfloor a^n \rfloor\}} = \frac{1}{\lfloor a^{N_x} \rfloor} \sum_{n=N_x}^{\infty} \frac{\lfloor a^{N_x} \rfloor}{\lfloor a^n \rfloor} \leq \frac{1}{x} \sum_{n=N_x}^{\infty} \frac{a^{N_x}}{a^n - 1},$$

since $\lfloor a^{N_x} \rfloor \geq x$ and $a^n \geq \lfloor a^n \rfloor \geq a^n - 1$. A simple calculation shows that

$$\sum_{n=N_x}^{\infty} \frac{a^{N_x}}{a^n - 1} \leq \sum_{n=0}^{\infty} \frac{1}{a^n - a^{-1}} < +\infty$$

for any N_x, which proves (3.13).

Now, using (3.12), applied at $\lfloor a^n \rfloor$ instead of n, and *Chebyshev's inequality*

$$\mathbb{P}(\bar{Y}_{\lfloor a^n \rfloor} - \bar{\mu}_{\lfloor a^n \rfloor} > \epsilon) \leq \frac{1}{\epsilon^2} \mathbb{E}[X_1^2 \frac{1}{\lfloor a^n \rfloor} \mathbf{1}_{\{|X_1| \leq \lfloor a^n \rfloor\}}].$$

Thus, from (3.13),

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{P}(|\bar{Y}_{\lfloor a^n \rfloor} - \bar{\mu}_{\lfloor a^n \rfloor}| > \epsilon) &\leq \frac{1}{\epsilon^2} \mathbb{E}[X_1^2 \sum_{n=1}^{\infty} \frac{1}{\lfloor a^n \rfloor} \mathbf{1}_{\{|X_1| \leq \lfloor a^n \rfloor\}}] \\ &\leq \frac{1}{\epsilon^2} \mathbb{E}[X_1^2 C_a |X_1|^{-1}] \\ &= \frac{C_a}{\epsilon^2} \mathbb{E}[|X_1|] < +\infty. \end{aligned}$$

The *Borel-Cantelli* then implies that $(a.s.) \lim \bar{Y}_{\lfloor a^n \rfloor} - \bar{\mu}_{\lfloor a^n \rfloor} = 0$, thus completing the proof of Lemma 3.3.9. □

3.4 Convergence of Infinite Series of i.i.d. and Applications

Let $\{X_n\}$ be an infinite sequence of random variables. Throughout S_n will denote the partial sum $\sum_{i=1}^{N} X_i$. We say that $\sum_{i=1}^{\infty} X_i$ converges almost surely if $\lim_{N \to \infty} S_N$ exists and is finite almost surely. We pose the question: under what conditions does $\sum_{i=1}^{\infty} X_i$ converge a.s.? An example of this problem is the random sign problem: if $\{c_n\}_1^{\infty}$ is a sequence of real numbers and if $\{\xi_n\}$ is a sequence of i.i.d. symmetric *Bernoulli* random variables ($\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = \frac{1}{2}$), what can we say about the convergence $\sum c_n \xi_n$? In fact there is a necessary and sufficient condition for the a.s. convergence of $\sum_{i=1}^{\infty} X_i$.

Theorem 3.4.1 *(Kolmogorov's Three Series Theorem)* Let $\{X_n\}$ be a sequence of independent random variables. Then $\sum_{i=1}^{\infty} X_i$ converges a.s. if and only if for some $a > 0$ all of the following three series converges:

(i) $\sum_{i=1}^{\infty} \mathbb{P}(|X_i| \geq a)$;
(ii) $\sum_{i=1}^{\infty} \mathbb{E}[X_i \mathbf{1}_{\{|X_i| \leq a\}}]$;
(iii) $\sum_{i=1}^{\infty} \mathrm{Var}(X_i \mathbf{1}_{\{|X_i| \leq a\}})$.

It is actually true that convergence of $\sum_{i=1}^{\infty} X_i$ implies the convergence of the three series in $(i)-(iii)$ for every $a>0$, but for the converse, convergence of $(i)-(iii)$ for any one $a>0$ suffices. We shall prove the sufficiency part of this theorem only. It is an easy consequence of the following special case, which is important enough to state separately as a theorem.

Theorem 3.4.2 If $\{X_n\}$ are independent random variables which all have 0 mean, then $\sum_{i=1}^{\infty} \mathrm{Var}(X_i) < +\infty$ implies that $\sum_{i=1}^{\infty} X_i$ converges a.s..

Notice that $\mathbb{E}[(S_N - S_M)^2] = \sum_{i=M+1}^{N} \mathrm{Var}(X_i)$ because $S_N - S_M = \sum_{i=M+1}^{N} X_i$ and the X_i are independent and have zero mean. The condition $\sum_{i=1}^{\infty} \mathrm{Var}(X_i) < +\infty$ then implies that $\lim_{M,N\to\infty} \mathbb{E}[(S_N - S_M)^2] = 0$, or, in words, that the sequence $\{S_n\}_1^{\infty}$ is *Cauchy* in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Hence, there is a random variable $Z \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[(S_n - Z)^2] \to 0$ as $n \to \infty$. Theorem 3.4.2 asserts that S_n converges to Z almost surely as well.

Theorem 3.4.2 resolves the random signs problem posed above. If $\xi_1, \xi_2, ...$ are independent, symmetric *Bernoulli* random variables, and if $c_1, c_2, ...$ is a sequence of real numbers, then

$$\sum_{i=1}^{\infty} \mathrm{Var}(c_i X_i) = \sum_{i=1}^{\infty} c_i^2 \tag{3.14}$$

and hence convergence of this sum is sufficient for almost sure convergence of $\sum_{i=1}^{\infty} c_i X_i$. The Theorem 3.4.1 implies that convergence of (3.14) is also necessary for almost sure convergence of the random sum. For a particular example, consider the sequence $c_n = \frac{1}{n}$. Since $\sum_{i=1}^{\infty} n^{-2} < +\infty$, we learnt that $\sum_{n=1}^{\infty} \xi_n/n$ converges almost surely, even though it is almost surely not *absolutely convergent*.

We shall prove Theorem 3.4.2 using an important inequality of *Kolmogorov* which is remarkable strengthening of *Chebyshev's inequality* for sums of independent random variables. *Kolmogorov's inequality* depends on the following simple observation concerning sum of random variables.

Lemma 3.4.3 Let $\{X_n\}$ be a sequence of independent, zero mean random variables with finite variances. If $m < n$ and if $A \in \sigma\{X_1, ..., X_m\}$ ($= \sigma$-algebra generated by $X_1, ..., X_m$), then

$$\mathbb{E}[S_m^2 \mathbf{1}_A] \le \mathbb{E}[S_n^2 \mathbf{1}_A]. \tag{3.15}$$

Proof. Observe that $S_n - S_m = \sum_{m+1}^{n} X_i$ is independent $X_1, ..., X_m$, and hence $\mathbb{E}[(S_n - S_m)S_m \mathbf{1}_A] = \mathbb{E}[S_n - S_m]\mathbb{E}[S_n - S_m]\mathbb{E}[S_m \mathbf{1}_A] = 0$. Thus

$$\begin{aligned}\mathbb{E}[S_n^2 \mathbf{1}_A] &= \mathbb{E}[(S_n - S_m)^2 \mathbf{1}_A] + 2\mathbb{E}[(S_n - S_m)S_m \mathbf{1}_A] + \mathbb{E}[S_m^2 \mathbf{1}_A] \\ &= \mathbb{E}[(S_n - S_m)^2 \mathbf{1}_A] + \mathbb{E}[S_m^2 \mathbf{1}_A],\end{aligned}$$

and inequality (3.15) follows. □

Proposition 3.4.4 *(Kolmogorov's inequality)* Let $\{X_n\}$ be a sequence of independent random variables with zero mean. Then

$$\mathbb{P}\left(\max_{1\leq n\leq N}|S_n|\leq\lambda\right)\leq\frac{\mathbb{E}[S_N^2]}{\lambda^2}.$$

Proof. Let A_k be the event $\{\,|S_n|<\lambda \text{ for } 1\leq n\leq k,\ |S_k|\geq\lambda\,\}$. In words, A_k is the event that the first time n such that $|S_n|$ rises above λ is $n=k$. The sets $A_1,...,A_N$ are disjoint and

$$\{\max_{1\leq n\leq N}|S_n|\geq\lambda\}=\bigcup_{k=1}^{N}A_k.$$

Now for each k, $1\leq k\leq N$, Lemma 3.4.3 and the *Markov inequality* imply

$$\mathbb{P}(A_k)\leq\lambda^{-2}\mathbb{E}[S_k^2\mathbf{1}_{A_k}]\leq\lambda^{-2}\mathbb{E}[S_N^2\mathbf{1}_{A_k}].$$

Thus

$$\mathbb{P}\left(\max_{1\leq n\leq N}|S_n|\geq\lambda\right)\leq\lambda^{-2}\sum_{k=1}^{N}\mathbb{E}[S_N^2\mathbf{1}_{A_k}]\leq\lambda^{-2}\mathbb{E}[S_N^2].$$

In the last step we used the disjointness of the A_k several times. □

Proof. (Theorem 3.4.2) Kolmogorov's inequality implies that

$$\mathbb{P}\left(\max_{M<k\leq N}|S_k-S_M|>\lambda\right)\leq\lambda^{-2}\mathbb{E}[(S_N-S_M)^2]=\lambda^{-2}\sum_{i=M+1}^{N}\mathrm{Var}(X_i).$$

Taking limits as $N\to\infty$,

$$\mathbb{P}\left(\sup_{M<k\leq\infty}|S_k-S_M|>\lambda\right)\leq\lambda^{-2}\sum_{i=M+1}^{\infty}\mathrm{Var}(X_i)$$

Because $\sum_{i=1}^{\infty}\mathrm{Var}(X_i)<+\infty$, we conclude that $\sup_{M<k\leq\infty}|S_k-S_M|$ converges to 0 in probability. It follows that $\sup_{M<k,M<l}|S_k-S_l|$ converges to 0 in probability. But any decreasing sequence of random variables that converges to 0 in probability must converge to 0 almost surely as well. Hence, $\sup_{M<k,M<l}|S_k-S_l|$ converges to 0 almost surely. This implies that the sequence $\{S_n\}_1^\infty$ is almost surely a *Cauchy* sequence. Hence, $\lim_{n\to\infty}S_n=\sum_{i=1}^{\infty}X_i$ converges almost surely. □

Proof. (Theorem 3.4.1) In the set up of Theorem 3.4.1, let $Y_i=X_i\mathbf{1}_{\{|X_i|>a\}}$ and let $\mu_i=\mathbb{E}[Y_i]$. Convergence of series (iii) implies by Theorem 3.4.1 that $\sum_{i=1}^{\infty}(Y_i-\mu_i)$ converges almost surely. Convergence of series (ii) says precisely that $\sum_{i=1}^{\infty}\mu_i$ converges, and it follows that $\sum_{i=1}^{\infty}Y_i$ converges a.s.. Now by convergence of series (i) and the *Borel-Cantelli lemma* $\mathbb{P}(X_i\neq Y_i\,i.o.\,)=0$ and hence convergence of $\sum_{i=1}^{\infty}Y_i$ implies that of $\sum_{i=1}^{\infty}X_i$. Since we have established a.s. convergence of $\sum_{i=1}^{\infty}Y_i$ a.s. convergence of $\sum_{i=1}^{\infty}X_i$ follows. □

We shall give several applications of Theorem 3.4.2 to large number law theorems. First we give a refinement of the second part of the *Borel-Cantelli lemma*. Then we give another proof of the strong law of large numbers, but under the slightly more restrictive hypothesis of mutual, rather than pairwise, independence of random variables. The applications depend on a simple lemma in the theory of infinite series.

Lemma 3.4.5 *(Kronecker's Lemma)* Let b_n be an increasing sequence of positive numbers such that $b_n \uparrow \infty$. Then if $\sum_{n=1}^{\infty} z_n/b_n$ converges to a finite limit,

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{i=1}^{b} z_i = 0.$$

Proof. (Sketch) A summation by parts shows that, with $f_k := \sum_{i=1}^{k} z_i/b_i$, $f_0 = 0$, $b_0 = 0$,

$$\frac{1}{b_n} \sum_{i=1}^{n} z_i = f_n - \frac{1}{b_n} \sum_{i=1}^{n} f_{i-1}(b_i - b_{i-1}). \tag{3.16}$$

Since $\lim_{n\to\infty} \frac{1}{b_n} \sum_{i=1}^{n} f_{i-1}(b_i - b_{i-1}) = \lim_{n\to\infty} f_n$, the right hand side of (3.16) tends to 0 as $n \to \infty$. □

We shall use *Kronecker's lemma* in conjunction with Theorem 3.4.2 to prove a strong version of the second part of the *Borel-Cantelli lemma*.

Theorem 3.4.6 Let $A_1, A_2, ...$ be independent events and let $p_i = \mathbb{P}(A_i)$, $1 \le i < +\infty$. If $\sum_{i=1}^{\infty} p_i = \infty$,

$$\frac{\sum_{i=1}^{n} \mathbf{1}_{A_i}}{\sum_{i=1}^{n} p_i} \to 1 \text{ almost surely as } n \to \infty. \tag{3.17}$$

Proof. Let

$$X_k = \frac{\mathbf{1}_{A_k} - p_k}{\sum_{i=1}^{k} p_i}.$$

Then the random variables $X_1, X_2, ...$ are independent, have zero mean, and

$$\sum_{k=1}^{\infty} \mathrm{Var}(X_k) = \sum_{k=1}^{\infty} \frac{p_k - p_k^2}{(\sum_{i=1}^{k} p_i)^2} \le \sum_{k=1}^{\infty} \frac{p_k}{(\sum_{i=1}^{k} p_i)^2}. \tag{3.18}$$

Note that for $k \ge 2$,

$$\frac{p_k}{(\sum_{i=1}^{k} p_i)^2} \le \frac{p_k}{\sum_{i=1}^{k} p_i \sum_{i=1}^{k-1} p_i} \le \frac{1}{\sum_{i=1}^{k} p_i} - \frac{1}{\sum_{i=1}^{k-1} p_i}.$$

Hence the last sum in (3.18) is bounded by the collapsing sum:

$$\sum_{k=1}^{\infty} \frac{1}{\sum_{i=1}^{k} p_i} - \frac{1}{\sum_{i=1}^{k-1} p_i} = 1 - p_1^{-1},$$

and this is finite. Therefore Theorem 3.4.2 applies to $\{X_n\}$, implying that $\sum_{i=1}^{\infty} X_i$ converges almost surely. But *Kronecker's lemma*, applied with $\mathbf{1}_{A_k} - p_k$ in the role of x_k and $\sum_{i=1}^{n} p_i$ in the role of b_n, then yields

$$\frac{\sum_{i=1}^{n} \mathbf{1}_{A_i} - p_i}{\sum_{i=1}^{n} p_i} \to 0, \text{ a.s. as } n \to \infty,$$

and (3.17) follows immediately. □

It is a simple corollary of the last result that when $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$, we have $\sum_{i=1}^{\infty} \mathbf{1}_{A_i} = \infty$ almost surely, and hence $\mathbb{P}(A_n \, i.o.\,) = 1$. Thus Theorem 3.4.6 includes the second half of the *Borel-Cantelli lemma*. Next, we shall apply *Kronecker's lemma* and Theorem 3.4.2 to prove the classical *Strong Law of Large Numbers* for an i.i.d. sequence of random variables with finite mean.

Theorem 3.4.7 Let $X_1, X_2, ...$ be i.i.d. random variables with $\mathbb{E}[|X_i|] < \infty$, $\mu := \mathbb{E}[X_i]$, then $(a.s.) \lim \bar{X}_n = \mu_X$.

Proof. We shall use again the truncation $Y_n := X_n \mathbf{1}_{\{|X_n| \le n\}}$ and its centering $Z_n = Y_n - \mathbb{E}[Y_n]$. As we know, we only need to prove $Z_n \to 0$ almost surely. Now *Kronecker's Lemma* says that $\bar{Z}_n \to 0$ almost surely if

$$\sum_{n=1}^{\infty} Z_n/n \text{ converges almost surely.} \tag{3.19}$$

On the other hand, Theorem 3.4.2 says that (3.19) is true if

$$\sum_{n=1}^{\infty} n^{-2} \mathrm{Var}(Z_n) < \infty. \tag{3.20}$$

We will show that this is true and hence complete the proof of Theorem 3.4.7. To verify (3.20), first observe that $\mathrm{Var}(Z_n) \le \mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \le n\}}]$ because the X_i's are identically distributed. Then notice that there is a constant C such that for any $x > 0$,

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \mathbf{1}_{\{x \le n\}} \le \frac{C}{x}.$$

Indeed, we see easily that

$$\sup_{x>0} x \sum n = 1^{\infty} \frac{1}{n^2} \mathbf{1}_{\{x \le n\}} < \infty$$

because for $x > 2$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \mathbf{1}_{\{x \leq n\}} \leq \int_{x-1}^{\infty} y^{-2} dy = (x-1)^{-1}.$$

Putting these observations together

$$\begin{aligned} \sum_{n=1}^{\infty} n^{-2} \mathrm{Var}(Z_n) &\leq \sum_{n=1}^{\infty} n^{-2} \mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \leq n\}}] \\ &= \mathbb{E}[X_1^2 \sum_{n=1}^{\infty} n^{-2} \mathbf{1}_{\{|X_1| \leq n\}}] \\ &\leq C\mathbb{E}[|X_1|] < +\infty \end{aligned}$$

This completes the proof. □

3.5 Stationary Process

Large number laws also hold for an important class of stochastic processes, called *stationary processes*, that greatly generalize *i.i.d.* sequences. Stationary processes appear in many applications as models of stochastic system whose dynamics are independent of time. In this section, we define and construct stationary processes; in the next, we state and prove the *ergodic theorem* for stationary processes, which addresses convergence of empirical means.

Recall that if $X_1, ..., X_n$ are random variables, $\mathbb{F}_{X_1,...,X_n}$ denotes their joint distribution measure; this is the probability measure defined on the *Borel sets* of $\mathbb{R}^n$ by

$$\mathbb{F}_{X_1,...,X_n}(U) = \mathbb{P}\left((X_1, ..., X_n) \in U\right).$$

Definition 3.5.1 A stochastic process $\mathbf{X} = \{X_n; n \geq 1\}$ is stationary if

$$\mathbb{F}_{X_1,...,X_m} = \mathbb{F}_{X_{n+1},...,X_{n+m}}, \quad \text{for all } n \geq 1 \text{ and } m \geq 1. \tag{3.21}$$

If the index n in $\{X_n\}$ is thought of as a time index, *stationarity* means that the statistical behaviour of the process is invariant with respect to shifts of time. It is sometimes useful to work with stochastic processes defined for all positive and negative integer times. In this case, the definition of stationarity is the same, except that the identity (3.21) must hold for all integers n.

If $\{X_n; n \geq 1\}$ is stationary, then, taking the case $m = 1$ in (3.21), $\mathbb{F}_{X_1} = \mathbb{F}_{X_{n+1}}$ for all n. Hence the random variables of a stationary process area always identically distributed.

Example 3.5.1 Let $\mathbf{X} = \{X_n; n \geq 1\}$ be an i.i.d. sequence, and let $\mathbb{F}$ denote the common distribution measure of the X_n's. Then, for any n,

$$\mathbb{F}_{X_{n+1},\ldots,X_{n+m}} = \mathbb{F} \times \cdots \times \mathbb{F} \quad (m \text{ times})$$

and hence $\mathbf{X}$ is stationary.

Example 3.5.2 Let $\{\xi_n; -\infty < n < \infty\}$ be a doubly infinite sequence of i.i.d. random variables. Assume $\mathbb{E}[\xi_n] = 0$ and $\mathrm{Var}(\xi_n) = \sigma^2$. Let $|a| < 1$, and define

$$X_n := \sum_{k=0}^{\infty} a^k \xi_{n-k}, \quad \text{for all } n, -\infty < n < \infty.$$

This sequence converges almost surely for each n, because the mean of each term, $a^k\xi_{n-k}$ is zero and $\sum_0^\infty \mathrm{Var}(a^k\xi_{n-k}) = \sum_0^\infty \mathrm{Var}\ a^{2k} = 1/(1-a^2) < +\infty$ (By *Kolmogorov's three series theorem*).

Such $\{X_n; -\infty < n < \infty\}$ is a *moving average*, using the weights a^k, $k \geq 0$, of the sequence $\{\xi_n; -\infty < n < \infty\}$. It is a simple and very basic model of a sequence of identically distributed, but uncorrelated random variables. A calculation shows that

$$X_n = aX_{n-1} + \xi_n. \tag{3.22}$$

One may think of this as the linear difference equation $x_n = ax_{n-1}$ perturbed by the noise sequence $\{\xi_n\}$. $\{X_n; -\infty < n < \infty\}$ may be thought of as the *"invariant"* solution obtained when the recursion defined by (3.22) has been allowed to run for an infinite amount of time. To make this precise, let $\{Y_n; n \geq -N\}$ solve $Y_n = aY_{n-1}+\xi_n$, for $n > -N$, with a given initial value $Y_{-N} = Z$. Then, an argument by induction shows that $Y_n = a^{n+N}Z + \sum_0^{n+N-1} a^k\xi_{n-k}$. Clearly, as $N \to \infty$, $Y_n \to X_n$, for each n.

To see that $\{X_n; -\infty < n < \infty\}$ is stationary, first write out the definition of X_n:

$$X_n = \xi_n + a\xi_{n-1} + a^2\xi_{n-2} + \cdots$$

The distribution of X_n is determined by the joint distribution of the sequence $\{\xi_n, \xi_{n-1}, ...\}$. Because $\{\xi_k; -\infty < k < \infty\}$ is an i.i.d. sequence, this joint distribution does not depend on n, and hence the distribution of X_n is not dependent on n. Now, for any $j \geq 1$,

$$X_{n+j} = a^j X_n + \sum_0^{j-1} a^k \xi_{n+j-k},$$

as we saw in the last paragraph. Notice that the second term depends only on $(\xi_{n+1}, ..., \xi_{n+j})$ which is independent of X_n, since X_n is defined as a function of $\{\xi_n, \xi_{n-1}, ...\}$. Thus,

$$\begin{pmatrix} X_{n+1} \\ X_{n+2} \\ X_{n+2} \\ \vdots \\ X_{n+m} \end{pmatrix} = \begin{pmatrix} a & 1 & 0 & 0 & 0 & 0 & \cdots \\ a^2 & a & 1 & 0 & 0 & 0 & \cdots \\ a^3 & a^2 & a & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ a^m & a^{m-1} & a^{m-2} & \cdots & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} X_n \\ \xi_{n+1} \\ \xi_{n+2} \\ \vdots \\ \xi_{n+m} \end{pmatrix}$$

Since the joint distribution of the random vector on the right-hand side of this equation does not depend on n, neither does the joint distribution of the random vector on the left, and hence $\{X_n; -\infty < n < \infty\}$ is stationary.

There is another way to generate stationary processes that is, in a sense, *canonical*. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A transformation $T : \Omega \mapsto \Omega$ is *measurable* if $T^{-1}(U) \in \mathcal{F}$ for every $U \in \mathcal{F}$. A measurable transformation is said to be *measure-preserving* on $(\Omega, \mathcal{F}, \mathbb{P})$ if $\mathbb{P}(T^{-1}(U)) = \mathbb{P}(U)$ for all $U \in \mathcal{F}$. Basic example of measure preserving transformations come from dynamical systems on compact spaces provided with a finite measure.

Example 3.5.3 For any real number a,

$$T(x) := (x + a) \bmod 1$$

is measure-preserving on $([0, 1), \mathcal{B}([0, 1)), \lambda)$, where λ denotes *Lebesgue measure*.

It will be useful to have the following equivalent characterization of the measure-preserving property.

Proposition 3.5.4 A *measurable transformation* T on $(\Omega, \mathcal{F}, \mathbb{P})$ is *measure-preserving* if and only if

$$\mathbb{E}[X] = \mathbb{E}[X \circ T] \tag{3.23}$$

for all integrable random variables X on $(\Omega, \mathcal{F}, \mathbb{P})$.

Proof. If A is an event, then $\mathbf{1}_A \circ T = \mathbf{1}_{T^{-1}(A)}$. If (3.23) is true for all integrable X, then for any event A,

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A] = \mathbb{E}[\mathbf{1}_A \circ T] = \mathbb{E}[\mathbf{1}_{T^{-1}(A)}] = \mathbb{P}(T^{-1}(A)),$$

and hence T is *measure-preserving*. Conversely, if T is measure-preserving, then (3.23) holds for all indicator functions. By linearity of expectation, it then holds for all simple random variables, and by taking limits, for all integrable random variables. □

The following theorem show how to generate stationary processes from measure-preserving transformations. In this theorem, T^k denotes the k-fold composition of T with itself, and $X \circ T^k$ denotes the random variable $X(T^k(\omega))$; T^0 denotes the *identity map*.

Theorem 3.5.5 Let T be a *measure-preserving transformation* on $(\Omega, \mathcal{F}, \mathbb{P})$. Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\{X_n := X \circ T^{n-1}; n \geq 1\}$ is a stationary process.

Proof. Fix an arbitrary $m \geq 1$ and an arbitrary Borel subset U of $\mathbb{R}^m$. Let

$$\begin{aligned} A :&= \{\omega; (X_1(\omega), X_2(\omega), ..., X_m(\omega)) \in U\} \\ &= \{\omega; \big(X(\omega), X(T(\omega)), ..., X(T^{m-1}(\omega))\big) \in U\} \end{aligned}$$

Then

$$\begin{aligned} T^{-1}(A) &= \{\omega; \left(X(T(\omega)), X(T^2(\omega)), ..., X(T^m(\omega))\right) \in U\} \\ &= \{\omega; (X_2(\omega), X_3(\omega), ..., X_{m+1}(\omega)) \in U\} \end{aligned}$$

Since T is measure-preserving, it follows that

$$\mathbb{P}\left((X_1, ..., X_m) \in U\right) = \mathbb{P}(A) = \mathbb{P}\left(T^{-1}(A)\right) = \mathbb{P}\left((X_2, ..., X_{m+1}) \in U\right).$$

By repeating this argument n times,

$$\mathbb{P}\left((X_1, ..., X_m) \in U\right) = \mathbb{P}(A) = \mathbb{P}(T^{-n}(A)) = \mathbb{P}\left((X_{n+1}, ..., X_{n+m}) \in U\right),$$

for any $n \geq 1$. Thus $\{X_n; n \geq 1\}$ is stationary. □

We shall now show that, at least at the level of distributions, any stationary process can be modeled using a measure-preserving transformation. Let $\mathbf{X} = \{X_n; n \geq 1\}$ be a stochastic processes on some probability space. We shall recall the definition of the distribution measure $\mathbb{F}_{\mathbf{X}}$ of the process. This is a probability space on the outcome space, $\mathbb{R}^\infty = \{x = (x_1, x_2, ...,) : x_i \in \mathbb{R} \text{ for all } i\}$, with the σ-algebra $\mathcal{B}(\mathbb{R}^\infty)$. Remember that $\mathcal{B}(\mathbb{R}^\infty)$ is generated by cylinder sets: that is,

$$\mathcal{B}(\mathbb{R}^\infty) = \sigma(\bigcup_n \pi_n^{-1}(\mathcal{B}(\mathbb{R}^n))),$$

where $\pi_n : \mathbb{R}^\infty \mapsto \mathbb{R}^n$ is the *projection* $\pi_n(x_1, ..., x_n, ...) = (x_1, ..., x_n)$. For any cylinder set $\pi_n^{-1}(B)$, where B is a Borel set of $\mathbb{R}^n$,

$$\{\omega; (X_1(\omega), X_2(\omega), ...) \in \pi_n^{-1}(B)\} = \{\omega; (X_1(\omega), ..., X_n(\omega)) \in B\}$$

is an event in $\mathcal{F}$. It follows that $\{\omega; (X_1(\omega), X_2(\omega), ...) \in U\} \in \mathcal{F}$ for every $U \in \mathcal{B}(\mathbb{R}^\infty)$, and thus that

$$\mathbb{F}_{\mathbb{X}}(U) = \mathbb{P}\left((X_1, X_2, ...) \in U\right), \quad U \in \mathcal{B}(\mathbb{R}^\infty).$$

makes sense and defines a probability measure on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$.

We will also need the canonical processes on $\mathbb{R}^\infty$. This is the process $\{Z_n; n \geq 1\}$, where $Z_n(x) = x_n$, for all $n \geq 1$ and all $x \in \mathbb{R}^\infty$. Under the probability measure $\mathbb{F}_{\mathbf{X}}$, the canonical process has the same distribution as the process $\mathbf{X} = \{X_n; n \geq 1\}$, because for any $U \in \mathcal{B}(\mathbb{R}^\infty)$,

$$\mathbb{P}\left((X_1, X_2, ...) \in U\right) = \mathbb{F}_{\mathbf{X}}(U) = \mathbb{F}_{\mathbf{X}}\left((Z_1, Z_2, ...) \in U\right).$$

Now, on $\mathbb{R}^\infty$ define the left-shift operator:

$$T_l(x_1, x_2, x_3, ...) = (x_2, x_3, ...).$$

T_l is measurable as a map from $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ to itself; indeed, if U is any set belonging to $\mathcal{B}(\mathbb{R}^\infty)$, then $T_l^{-1}(U) = \mathbb{R} \times U$, which is again a set in $\mathcal{B}(\mathbb{R}^\infty)$. Observe that the canonical process can be defined by composing the left-shift with Z_1l for all $n \geq 1$,

$$\left[Z_1 \circ T^{n-1}\right](x_1, x_2, ...) = Z_1(x_n, x_{n=1}, ...) = x_n = Z_n(x).$$

Theorem 3.5.6 The following statements are equivalent:

- $\mathbf{X} := \{X_n; n \geq 1\}$ is stationary;
- T_l is measure preserving on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{F}_\mathbb{X})$;
- The canonical process $\{Z_n = Z_1 \circ T^{n-1}; n \geq 1\}$ on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{F}_\mathbf{X})$ is stationary.

Proof. The equivalence of (2) and (3) is immediate from chasing through the definitions and applying Theorem 3.5.5. (3) implies (1) because for any $n \geq 0$, $m \geq 1$, and $U \in \mathcal{B}(\mathbb{R}^m)$,

$$\begin{aligned}\mathbb{P}\left((X_{n+1}, ..., X_{n+m}) \in U\right) &= \mathbb{F}_\mathbf{X}\left((Z_{n+1}, ..., Z_{n+m}) \in U\right) \\ &= \mathbb{F}_\mathbf{X}\left((Z_1, ..., Z_m) \in U\right) \\ &= \mathbb{P}\left((x_1, ..., X_m) \in U\right).\end{aligned}$$

These same equations for $n = 1$ shows that, if (1) is true, when $\mathbb{F}_\mathbf{X}(T_l^{-1}(U)) = \mathbb{F}_\mathbf{X}(U)$ for every cylinder set $U \subset \mathcal{F}(\mathbb{R}^\infty)$. But two probability measures that agree on an algebra, agree on the σ-algebra generated by the algebra. Since $\mathcal{B}(\mathbb{R}^\infty)$ is generated by the algebra of cylinder sets, it follows that $\mathbb{F}_\mathbf{X} \circ T_l^{-1} = \mathbb{F}_\mathbf{X}$. Thus (1) implies (2). □

Because of this theorem, any question about the distribution of a stationary process can be phrased in terms of a stationary process defined by a random variable and a measure preserving transformation on a probability space. The final result of this section shows how to define new stationary processes from old ones.

Theorem 3.5.7 Let $\{X_n; -\infty < n < \infty\}$ be a stationary process, and let $\phi : \mathbb{R}^\infty \mapsto \mathbb{R}$ be any $\mathcal{B}(\mathbb{R}^\infty)$-measurable function. Then

$$Y_k := \phi(X_k, X_{k-1}, X_{k-2}, ...), \quad k \geq 1, \quad \textit{is stationary.}$$

Observe that the *moving average* process is a consequence of this theorem.

3.6 The Ergodic Theorem

Throughout this section, $X_n := X \circ T^{n-1}$, $n \geq 1$, where X is a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and T is a measure-preserving transformation of Ω. Since

$X_k(\omega) = X(T^{k-1}(\omega))$, the sequence $X_1(\omega), X_2(\omega), ..., X_n(\omega), ...$ represents the values of X along the *orbit*, $(\omega, T(\omega), T^2(\omega), ...)$ of T. As we shall see in this section, the measure preserving property of T is enough to ensure that the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{1}^{n} X_k(\omega)$$

exists almost surely. In addition, it is possible to uniquely characterize this limit as a random variable. This characterization leads to condition called *ergodicity*, under which the limit is almost-surely equal to $\mathbb{E}[X]$, so that the law of large number holds. As in the discussion of the strong law of large numbers, $\bar{X}_n$ shall sometimes be used to denote $\frac{1}{n}\sum_1^n X_k$. To state the main result, called the *ergodic theorem*, we need the concepts of invariant events and invariant random variables. To motivate the definitions, observe that

$$\begin{aligned} X_n(\omega) :&= \frac{1}{n} \sum_{1}^{n} X(T^{k-1}(\omega)) \\ &= \frac{X(\omega)}{n} + \frac{n-1}{n}\left[\frac{1}{n-1}\sum_{1}^{n-1} X\left(T^{j-1}(T(\omega))\right)\right] \\ &= \frac{X(\omega)}{n} + \frac{n-1}{n}\bar{X}_{n-1}(T(\omega)) \end{aligned}$$

Therefore, if we let $V := \{\omega; \lim_{n\to\infty} \bar{X}_n(\omega) \text{ exists}\}$, it follows that $\omega \in V$ if and only if $T(\omega) \in V$, or equivalently, $T^{-1}(V) = V$. Furthermore, if $\omega \in V$, it also follows that $\lim_{n\to\infty} \bar{X}_n(\omega) = \lim_{n\to\infty} \bar{X}_n(T(\omega))$; equivalently, if we define $Y(\omega) = \mathbf{1}_V(\omega) \lim_{n\to\infty} \bar{X}_n(\omega)$, then $Y(\omega) = Y(T(\omega))$.

Definition 3.6.1 An event $A \in \mathcal{F}$ is *invariant* with respect to T if $T^{-1}(A) = A$, or equivalently, if $\mathbf{1}_A = \mathbf{1}_A \circ T$. A random variable is *invariant* with respect to T if $X = X \circ T$.

When the transformation T is clear from context, we shall just refer to invariant sets or invariant random variables without specifically mentioning T. The collection of all invariant events shall be denoted by $\mathcal{I}$. The following basic facts are left to the reader to prove: (1) $\mathcal{I}$ is a σ-algebra; and, (2) X is invariant if and only if X is $\mathcal{I}$-measurable.

Now follows one of the most important definitions of the theory:

Definition 3.6.2 A *measure-preserving transformation* T on $(\Omega, \mathcal{F}, \mathbb{P})$ is called *ergodic*, if the invariant σ-algebra is trivial; that is, $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ for every invariant event A.

In general, checking whether or not T is ergodic can be tricky. We shall state two examples here, but we defer the proofs to a later discussion:

- The left shift T_l on $\mathbb{R}^\infty$ is ergodic with respect to any product probability measure $\mathbb{F}^\infty$; $(\mathbb{F}^\infty := \prod_1^\infty \mathbb{F})$;

- Let α be an irrational number. Then $T(x) = [x+\alpha] | \mod 1$ is ergodic on $([0,1), \mathcal{B}([0,1)), \lambda)$.

We are now ready to state the ergodic theorem for measure-preserving transformations.

Theorem 3.6.1 Let T be a *measure-preserving transformation* on $(\Omega, \mathcal{F}, \mathbb{P})$. Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying $\mathbb{E}[|X|] < +\infty$. Then there is an invariant random variable Y such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\infty} X \circ T^{k-1} = Y, \text{ a.s..} \tag{3.24}$$

The convergence holds in L^1 as well, that is,

$$\lim_{n\to\infty} \mathbb{E}\left[|\frac{1}{n} \sum_{k=1}^{n} X \circ T^{k-1} - Y| \right] = 0. \tag{3.25}$$

The limit Y is unique (up to a.s. equivalence) random variable for which:

1. Z is $\mathcal{I}$-measurable; and
2. For every $A \in \mathcal{I}$, $\mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}[\mathbf{1}_A X]$.

When T is *ergodic*, $Y = \mu$ almost surely, that is,

$$\frac{1}{n} \sum_{k=1}^{n} X \circ T^{k-1} \text{ converges to } \mu \text{ a.s. and in } L^1. \tag{3.26}$$

An interesting feature of this theorem is that it gives a measure theoretic characterization of the limit. We shall see later that Y is the conditional expectation of X given the invariant σ-algebra $\mathcal{I}$. What this means heuristically is that the limit Y is a kind of average of X over the invariant sets. This idea is contained in the conditions stated in the theorem that Y is measurable with respect to the invariant σ-algebra and that $\mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}[\mathbf{1}_A X]$ for every invariant A. For example, suppose that A is an invariant set such that $\mathbb{P}(A) > 0$ and there is no invariant subset B of A, with $0 < \mathbb{P}(B) < \mathbb{P}(A)$. Then there is a real number b such that $Y = b$ almost surely on A, for if not, one could find a value c such that $A \cap \{Y < c\}$ and $A \cap \{Y \geq c\}$ are both events with positive probability, and then both of these subsets would be invariant subsets of A with positive probabilities strictly less than $\mathbb{P}(A)$. For this value of b, it would then follow that $b\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}[\mathbf{1}_A X]$, or $Y = b = \mathbb{E}[\mathbf{1}_A X]/\mathbb{P}(A)$, which says that Y is an average of X over A. It could happen there may be no invariant A of the type assumed here, but this argument still conveys the essential idea.

This argument of the previous paragraph also explains the definition of ergodicity and why (3.24) and (3.25) imply the result in (3.26) that the empirical means $\bar{X}_n$ converge to $\mathbb{E}[X]$

as $n \to \infty$. When T is *ergodic*, Ω contains no invariant event A such that $0 < \mathbb{P}(A) < 1$. By the argument above, invariant random variable is then almost surely constant on Ω. Thus, there is a constant b such that $b = Y = \lim_{n\to\infty} \bar{X}_n$ almost surely. It follows then that $Y = b = \mathbb{E}[\mathbf{1}_\Omega Y] = \mathbb{E}[\mathbf{1}_\Omega] = \mathbb{E}[X]$, as we wanted to show.

Proof of the ergodic theorem:
The proof of the ergodic theorem relies on the maximal ergodic inequality, which we state and prove next. This result looks simple, but is, in fact, deep, as we shall see.

Throughout, let $S_0 := 0$, and for positive integers n, introduce the notation,

$$S_n := \sum_{k=1}^{n} X \circ T^{k-1}, \quad M_n := \sup_{1\le k\le n} S_k, \quad M := \sup_{1\le k<\infty} S_k.$$

The ergodic maximal inequality exploits the additive structure of the process $\{S_n\}$, as expressed in the following identity, whose proof is an easy calculation:

$$S_{n+1}(\omega) = X(\omega) + S_n(T(\omega)), \quad n \ge 0 \tag{3.27}$$

As a consequence,

$$M_{n+1}(\omega) = X(\omega) + \max_{0\le k\le n} S_k(T(\omega)) = X(\omega) + \max\{0, M_n(T(\omega))\} \tag{3.28}$$

This is the principal observation.

Lemma 3.6.2 For any invariant event A,

$$\mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M>0\}} X\Big] \ge 0. \tag{3.29}$$

Proof. Since A is invariant $\mathbf{1}_A = \mathbf{1}_A \circ T$. Equality (3.28) implies:

$$\mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} X\Big] = \mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} M_{n+1}\Big] - \mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} \max\{0, M_n \circ T\}\Big]. \tag{3.30}$$

However, $\max\{0, M_n \circ T\} = \mathbf{1}_{\{M_n\circ T>0\}} M_n \circ T$, and so

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} \max\{0, M_n \circ T\}\Big] &\le \mathbb{E}\Big[\mathbf{1}_A \max\{0, M_n \circ T\}\Big] \\
&= \mathbb{E}\Big[(\mathbf{1}_A \circ T)\mathbf{1}_{\{M_n\circ T>0\}} M_n \circ T\Big] \\
&= \mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} M_n\Big],
\end{aligned}$$

The last equality used the measure-preserving property of T. Using this in (3.30) gives

$$\mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} X\Big] \geq \mathbb{E}\Big[\mathbf{1}_A \mathbf{1}_{\{M_n>0\}} (M_{n+1} - M_n)\Big] \geq 0. \tag{3.31}$$

Notice that the sequence of random variables $\mathbf{1}_{\{M_n>0\}}$, $n \geq 1$, is increasing and converges everywhere to $\mathbf{1}_{\{M>0\}}$. Thus inequality (3.29) follows from (3.31) by letting $n \to \infty$ and using dominated convergence theorem. □

The next step is to use the maximal ergodic inequality to prove the existence of the limit $Y = \lim_{n\to\infty} S_n/n$. For every real number b, define the events:

$$U_b := \left\{ \limsup \frac{S_n}{n} > b \right\} \text{ and } L_b := \left\{ \liminf \frac{S_n}{n} < b \right\}.$$

Notice that U_b an L_b are invariant events.

Let B be any invariant event. Now apply maximal ergodic inequality (3.29), with $B \cap U_b$ in place of A and $X - b$ in place of X. When these substitutions are made, S_n should be replaced by $\sum_1^n (X-b) \circ T^{k-1} = S_n - nb$ and M by $\sup_n S_n - nb$. The result is:

$$\mathbb{E}\Big[\mathbf{1}_B \mathbf{1}_{U_b} \mathbf{1}_{\{\sup_n (S_n - nb)>0\}} (X-b)\Big] \geq 0. \tag{3.32}$$

However, $U_b = \{\limsup S_n/n > b\} \subset \{\sup_n (S_n - nb) > 0\}$, and hence $\mathbf{1}_{U_b} \mathbf{1}_{\{\sup_n (S_n-nb)>0\}} = \mathbf{1}_{U_b}$. Thus, from (3.32),

$$\mathbb{E}\Big[\mathbf{1}_B \mathbf{1}_{U_b} (X-b)\Big] \geq 0 \tag{3.33}$$

Similarly, by applying the maximal ergodic inequality to $a - X$,

$$\mathbb{E}\Big[\mathbf{1}_B \mathbf{1}_{L_a} (X-a)\Big] \leq 0, \tag{3.34}$$

for any invariant B and any a. From these inequalities we can conclude the following,

Lemma 3.6.3 For any $a < b$,

$$\mathbb{P}\left(\liminf \frac{S_n}{n} < a < b < \limsup \frac{S_n}{n}\right) = \mathbb{P}\left(L_a \cap U_b\right) = 0. \tag{3.35}$$

Also,

$$\mathbb{P}\left(\liminf \frac{S_n}{n} = -\infty\right) = 0 \text{ and } \mathbb{P}\left(\limsup \frac{S_n}{n} = \infty\right) = 0 \tag{3.36}$$

Proof. Notice that as b increases, the events U_b decrease down to the event $\{\limsup \frac{S_n}{n} = \infty\}$ and so $\mathbb{P}(\limsup \frac{S_n}{n} = \infty) = \lim_{b\to\infty} \mathbb{P}(U_b)$. Now apply (3.33) with $B = \Omega$ and rearrange terms. The result is:

$$b\mathbb{P}(U_b) = b\mathbb{E}\big[\mathbf{1}_{U_b}\big] \le \mathbb{E}\big[\mathbf{1}_{U_b} X\big] \le \mathbb{E}\big[|X|\big].$$

Hence, $\mathbb{P}(U_b) \le \mathbb{E}\big[|X|\big]/b \to 0$ and this tends to 0 as $b \to \infty$, which completes the proof of the second equality in (3.36). The first equality is proved by a similar argument using (3.34).

As for (3.35), let $a < b$, apply (3.33) and (3.34) with $B = U_a \cap L_b$. After rearrangement of terms, the result is:

$$b\mathbb{P}(U_a \cap L_b) \le \mathbb{E}\big[\mathbf{1}_{U_a \cap L_b} X\big] \le a\mathbb{P}(U_a \cap L_b).$$

Since $a < b$, this can only be true if $\mathbb{P}(U_a \cap L_b) = 0$, which proves (3.35). □

We are now ready to finish the proof of the ergodic theorem. The event that S_n/n does not converge to a finite limit is contained in

$$\left\{ \liminf \frac{S_n}{n} = -\infty \right\} \bigcup \left\{ \limsup \frac{S_n}{n} = \infty \right\} \bigcup \left[\bigcup_{a<b, a,b\in\mathbb{Q}} L_a \cap U_b \right],$$

where $\mathbb{Q}$ denotes the rational numbers. But it follows immediately from previous lemma that this event has probability zero. Thus, $\frac{S_n}{n}$ converges with probability one. Set $Y(\omega) := \lim S_n(\omega)/n$, if the limit exists, and $Y(\omega) = 0$, otherwise. Then Y is invariant and $S_n/n \to Y$ almost surely, proving (3.24) of the ergodic theorem.

We will show next that S_n/n converges to Y in L^1-norm. Suppose first that $|X|$ is bounded, say by a constant K. Then $|S_n/n|$ and hence Y are also bounded by K and it follows by *dominated convergence* that $\mathbb{E}\big[|S_n/n - Y|\big] \to 0$ as $n \to \infty$. This establishes L^1-convergence for the bounded case.

Now assume only that $\mathbb{E}[|X|] < +\infty$, and let $Y = \lim_n S_n/n$. For any $K > 0$, let $X^k := X\mathbf{1}_{\{|X|\le K\}}$, $S_n^K = \sum_1^n X^K \circ T^{k-1}$, and

$$Y^K := \lim_n \frac{1}{n} \sum_1^n X^k \circ T^{k-1},$$

which we know exists. By *Fatou's lemma*, the triangle inequality and characterization of measure-preserving transformation,

$$\mathbb{E}\bigg[|Y - Y^k|\bigg] \le \liminf_n \mathbb{E}\bigg[|\frac{1}{n}\sum_1^n (X - X^k) \circ T^{k-1}|\bigg] \le \mathbb{E}\bigg[|X - X^k|\bigg].$$

Therefore,

$$\mathbb{E}\left[|\frac{1}{n}S_n - Y|\right] \leq \mathbb{E}\left[|S_n - S_n^K|\right] + \mathbb{E}\left[|\frac{1}{n}S_n^K - Y^K|\right] + \mathbb{E}\left[|Y - Y^K|\right]$$
$$\leq 2\mathbb{E}\left[|X - X^K|\right] + \mathbb{E}\left[|\frac{1}{n}S_n^K - Y^K|\right]$$

Take $n \to \infty$, then the last term goes to 0 since X^K is bounded, and thus,

$$\limsup_n \mathbb{E}\left[|\frac{1}{n}S_n - Y|\right] \leq 2\mathbb{E}\left[|X - X^K|\right].$$

But this is true for any positive K. Thus, taking $K \to \infty$ and use dominated convergence theorem to obtain:

$$\limsup_n \mathbb{E}\left[|\frac{1}{n}S_n - Y|\right] = 0$$

This proves (3.25) of the statement of ergodic theorem.

Finally, it remains to prove that Y is unique invariant random variable such that

$$\mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}[\mathbf{1}_A X] \text{ for all invariant } A. \tag{3.37}$$

For any invariant set A, $A = T^{-n+1}A$ for any positive integer n, and thus,

$$\mathbb{E}\left[\mathbf{1}_A X \circ T^{n-1}\right] = \mathbb{E}\left[\left(\mathbf{1}_A \circ T^{n-1}\right) X \circ T^{n-1}\right] = \mathbb{E}[\mathbf{1}_A X].$$

As a result,

$$\mathbb{E}\left[\mathbf{1}_A \frac{1}{n}\sum_1^n X \circ T^{k-1}\right] = \mathbb{E}[\mathbf{1}_A X],$$

for every positive integer n. Taking limits as $n \to \infty$ and using the L^1-convergence stated in (3.25),

$$\mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}[\mathbf{1}_A X].$$

Now suppose Y' is another invariant random variable satisfying (3.37). The event $\{Y > Y'\}$ is invariant since both Y and Y' are invariant. Applying property (3.37) for both Y and Y',

$$\mathbb{E}\left[\mathbf{1}_{\{Y>Y'\}} Y\right] = \mathbb{E}\left[\mathbf{1}_{\{Y>Y'\}} X\right] = \mathbb{E}\left[\mathbf{1}_{\{Y>Y'\}} Y'\right],$$

and hence,

$$\mathbb{E}\left[\mathbf{1}_{\{Y>Y'\}}(Y - Y')\right] \geq 0$$

It follows that $\mathbb{P}(Y > Y') = 0$. Reversing the roles of Y and Y' gives also $\mathbb{P}(Y < Y') = 0$, and so $\mathbb{P}(Y = Y') = 1$. Finally, we have already shown in the discussion after the statement of the ergodic theorem that $Y = \mathbb{E}[X]$ almost surely when T is ergodic. This completes the proof of the ergodic theorem.

The rest of this section will be devoted to proving ergodicity of the two examples stated above.

Theorem 3.6.4 The left shift is ergodic on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{F}^\infty)$.

Proof. Let $Z_1(x) = x_1$ for $x = (x_1, x_2, ...) \in \mathbb{R}^\infty$. Then $Z_n = Z_1 \circ T^{n-1}$ defines the canonical process, and, on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{F}^\infty)$, $\{Z_n; n \geq 1\}$ is a sequence of i.i.d. random variables with common distribution $\mathbb{F}$. Let $\mathcal{T} = \cap_n \sigma(Z_n, Z_{n+1}, ...)$ be the tail σ-algebra. *Kolmogorov's zero-one law* says that $\mathcal{T}$ is trivial – for every event $A \in \mathcal{T}$, either has measure 0 or 1. Let $\mathcal{I}$ be the set of events invariant under the left shift. We will show that $I \subset \mathcal{T}$, which will complete the proof. Let B be invariant; then $T^{-n}B = B$ for all integers $n \geq 1$. Write B as $B = \{x; (Z_1(x), Z_2(x), ...) \in B\}$. Then for every n, $B = T^{-n}B = \{x; (Z_{n+1}, Z_{n+2}, ...) \in B\}$. Thus $B \in \sigma(Z_{n+1}, Z_{n+2}, ...)$ for all n, and so B is in the tail σ-algebra. □

Theorem 3.6.5 Let α be an irrational number. Then $T(x) = [x + \alpha]|$ mod 1 is ergodic on $([0, 1), \mathcal{B}([0, 1)), \lambda)$.

Proof. Let $X(x)$ define a bounded invariant random variable. By the theory of *Fourier series*, $\{e^{i2\pi nx}; -\infty < n < \infty\}$ is a complete orthonoraml basis of the space $L^2([0, 1), dx)$ of square-integrable functions on $[0, 1)$. Thus, the infinite series

$$\sum_{-\infty}^{\infty} a_n e^{2\pi inx}, \text{ where } a_n = \int_0^1 X(x)e^{-2\pi inx}dx$$

converge in L^2 to X. But, because X is invariant,

$$\begin{aligned} a_n = \int_0^1 X(x)e^{-2\pi inx}dx &= \int_0^1 X(x+\alpha)e^{-2\pi inx}dx \\ &= \int_0^1 X(x)e^{-i2\pi n(x-\alpha)}dx = e^{i2\pi n\alpha}a_n \end{aligned}$$

If $n \neq 0$ and α is irrational, this can only be true if $a_n = 0$. It follows that

$$X(x) = a_0 = \int_0^1 X(x)dx$$

almost everywhere. If $X(x) = \mathbf{1}_A(x)$ where A is an invariant set for T, it follows that $\mathbf{1}_A(x) = \lambda(A)$ for almost everywhere X. Therefore, either $\lambda(A) = 1$. Thus, T is *ergodic*. □

3.7 The ergodic theorem for general stationary processes

In the previous section, we state and proved the ergodic theorem for stationary processes that are defined using a measure preserving transformation. In this section, we will translate this into a theorem for stationary process, generally defined.

Let $\mathbf{X} = \{X_n; n \geq 1\}$ be a stationary processes defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathbb{F}_{\mathbf{X}}$ be its distribution measure on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$. We know then that the left shift is a measure preserving transformation on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathbb{F}_{\mathbf{X}})$.

We want to define invariant event and ergodicity directly for the process $\mathbf{X}$ on its original probability space. We do this as follows. An event A is invariant for $\mathbf{X}$ if there exists a subset $B \in \mathcal{B}(\mathbb{R}^\infty)$ such that

$$A = \{\omega : (X_n(\omega), X_{n+1}(\omega), ...) \in B\} = \{\omega; T_l^{n-1}(X_1(\omega), X_2(\omega), ...) \in B\}$$

for all positive integers n. When this is the case,

$$A = \{\omega; (X_1(\omega), X_2(\omega), ...) \in T_l^{-n}(B)\} \text{ for all } n$$

It follows that $\mathbb{F}_{\mathbf{X}}(B \Delta T_l^{-1}(B)) = 0$, since $\{\omega; X_1(\omega), X_2(\omega), ...\} \in B \Delta T_l^{-1}(B)$ is empty. We say in this case that B is quasi-invariant with respect to T_l and $\mathbb{F}_{\mathbf{X}}$. If B is invariant with respect to T_l, then it is clear that $\{\omega; (X_1(\omega), X_2(\omega), ...) \in B\}$ is invariant. The class of invariant events A for $\mathbf{X}$ will be denoted by $\mathcal{I}(\mathbf{X})$; it is a σ-algebra, a fact which is easily checked. We say that $\mathbf{X}$ is ergodic if $\mathcal{I}(\mathbf{X})$ is trivial. It is clear from the discussion that $\mathbf{X}$ is ergodic implies T_l is ergodic with respect to $\mathbb{F}_{\mathbf{X}}$.

Now we can restate the ergodic theorem for stationary process:

Theorem 3.7.1 Let $\mathbf{X} = \{X_n\}$ be a stationary process such that $\mathbb{E}[|X_1|] < +\infty$. Then there is an $\mathcal{I}(\mathbf{X})$-measurable random variable Y such that $\frac{1}{n}\sum_1^n X_k$ converges to Y almost surely and in $L^1(\mathbb{P})$. Moreover, Y is uniquely characterized by the conditions:

1. Z is $\mathcal{I}(\mathbf{X})$-measurable;
2. For every $A \in \mathcal{I}(\mathbf{X})$, $\mathbb{E}[\mathbf{1})_A(Y)] = \mathbb{E}[\mathbf{1}_A X_1]$.

Finally, if $\mathbf{X}$ is ergodic, then $Y = \mathbb{E}[X_1]$.

Also as an extension of previous theorem,

Theorem 3.7.2 Let $\{X_n; n \geq 1\}$ be an ergodic stationary process, and let $\phi : \mathbb{R}^\infty \mapsto \mathbb{R}$ be any $\mathcal{B}(\mathbb{R}^\infty)$-measurable function. Then

$$Y_k := \phi(X_k, X_{k-1}, X_{k-2}, ...), \ k \geq 1, \text{ is ergodic.}$$

Chapter 4

Convergence in Distribution and the Central Limit Theorem

Let $X_1, X_2, ...$ be i.i.d. random values with finite mean, μ, and finite variance, σ^2, and let $S_n = \sum_1^n X_i$, $n \geq 1$, denote the associated partial sum process. Law of large number theorems tell us that the empirical mean, S_n/n, converges to the true mean μ in mean square and almost surely, as $n \to \infty$. These results could be thought of as a leading order asymptotic analysis of the process $\{S_n/n\}$. It is natural to ask if an asymptotic analysis be pursued to the next order. Namely, what can be said about the fluctuation, $S_n/n - \mu$, of the empirical mean about μ, as $n \to \infty$? How fast does it approach zero, and what statistical structure does it have, if any? Can one write $S_n/n = \mu + Z_n/a_n$ where $a_n \to \infty$ and Z_n is converging in some sense, but not to zero? This question, and its extension to more general sequence $X_1, X_2, ...$, are a major theme of probability theory. The answer is 'yes, in great generality', and the main results have the flavour of what physicists call *universality*: they are independent of the distribution of the common probability law of the X_n's. For this reason, these results are called *central limit theorem.*

The correct scale at which to study $S_n/n - \mu$ is easy to guess. Its variance is $n\sigma^2$, and therefore, if

$$Z_n := \frac{\sqrt{n}}{\sigma}\big[\frac{S_n}{n} - \mu\big] = \frac{(S_n - n\mu)}{\sigma\sqrt{n}}. \text{ or equivalently, } \frac{S_n}{n} = \mu + \frac{\sigma}{\sqrt{n}}Z_n,$$

the variance of Z_n equals one for every n. Hence, Z_n is not converging to 0 in probability, and, at least in a distributional sense, the size of $S_n/n - \mu$ is on the order of $\frac{1}{\sqrt{n}}$. What about Z_n? The classical theorem of central limit theory is the following remarkable result.

Theorem 4.0.3 If $X_1, X_2, ...$ is a sequence of i.i.d. random variables with mean μ and variance σ^2, then

$$\lim_{n\to\infty} \mathbb{P}\left(a < Z_n \leq b\right) = \int_a^b e^{-x^2/2}\frac{dz}{\sqrt{2\pi}} \tag{4.1}$$

The function under the integral sign on the right-hand side of (4.1) is the probability density of a standard normal random variable, a normal random variable with zero mean and unit variance, and so this right hand side is the probability a standard normal takes values between a and b. In fact, the theorem is saying that, as $n \to \infty$, Z_n 'looks' more and more like a standard normal. The concept of convergence in distribution, defined below, will make this statement more precise. No assumption is made on the common distribution of $X_1, X_2, ...$ beyond the existence of a finite mean and variance; the normal distribution is universal limit law for the scaled fluctuations of the empirical mean about its true mean. There is no importance to choice of strict and non-strict inequalities in $a < Z_n \leq b$, or $a \leq Z_n < b$. We choose this combination because $\mathbb{P}(a < Z_n \leq b) = F_{Z_n}(b) - F_{Z_n}(a)$, so that the statement translates immediately to a claim about the limit of the cumulative distribution function of Z_n.

Theorem 4.0.3 suggests many other questions. For example, if n is large and $X_1, X_2, ..., X_n$ are independent and identically distributed and $\text{Var}(X_1 + \cdots + X_n) = 1$, what we can say approximately about the distribution of $X_1 + \cdots + X_n$. Of course, the rigorous formulation of this question is about a limit as $n \to \infty$. What kinds of limiting distribution arise? Under what conditions are these limiting distribution normal? What if X_1, X_2, ... are not identically distributed? The *Linderberg-Feller CLT*, stated and proved later in the chapter, states roughly that as long as all X_i, $i \leq n$, have roughly the same influence on the sum S_n, for all n, then there is a normal limit. This result is one reason for the popularity and appropriateness of the normal distribution in many statistical models. A random outcome, built from the influence of many small random contributions of about the same strength, should look approximately normal.

4.1 DeMoivre-Laplace Central Limit Theorem

Historically, the *Central Limit Theorem (CLT)* goes back to *DeMoivre*, who in 1738 established the normal limit in (4.1) when $X_1, X_2, ...$ are i.i.d. *Bernoulli* random variables. This case can be derived by explicit calculations, which are very instructive to carry out. We present one derivation in this section. In fact, we prove a stronger statement in *Theorem 4.1.1* below called a *local central limit theorem*. What this means precisely, will be clear from the statement of the result.

Throughout this section, X_1, X_2,... are i.i.d. *Bernoulli* random variables. Thus, $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q := 1 - p$, for all i. As usual, $S_n = \sum_1^n X_i$ denotes the sum. As we know, S_n is a binomial(n, p) random variable:

$$\mathbb{P}(S_n = m) = C_n^m p^m q^{n-m}, \ m = 1, 2, ..., n.$$

Since $\mathbb{E}[S_n] = np$ and $\text{Var}(S_n) = npq$, the scaled fluctuation of the empirical mean about p

is thus,

$$Z_n := \frac{npq}{\sqrt{n}}\big[\frac{S_n}{n} - p\big] = \frac{S_n - np}{\sqrt{npq}}$$

For every n, Z_n is a discrete random variable, and we are trying to compare it to the standard normal, which is a continuous random variable with a density. To build a bridge between discrete and continuous, we will define a density function that generates the distribution approximating that of Z_n. The probability that a continuous random variable with density f lies between a and b is the area $\int_a^b f(x)dx$ under f from $x = a$ to $x = b$. So we will proceed by defining a curve from which probabilities of Z_n can be approximated by computing areas.

Let

$$x_m^n := \frac{m - np}{\sqrt{npq}}$$

Thus,

$$x_0^n = \frac{-np}{\sqrt{npq}},\ x_1^n = \frac{1 - np}{\sqrt{npq}},\ ...,\ x_n^n = \frac{n(1-p)}{\sqrt{npq}},$$

are the possible values of Z_n. These points define a partition of $[x_0^n, x_n^n]$ into equal sized sub-intervals of length $\frac{1}{\sqrt{npq}}$. Now, above each point x_m^n, $1 \leq m \leq n-1$ construct a rectangle whose base extends between the midpoints of the sub-intervals on either side of x_m^n and whose area is the probability that Z_n equals to x_m^n; hence, since the base of this rectangle is of length $1/\sqrt{npq}$, its height is $\sqrt{npq}C_n^m p^m q^{n-m}$. Construct rectangles of the same width and over the endpoints x_0^n and x_n^n as well, whose respective areas are $\mathbb{P}(Z_n = x_0^n)$ and $\mathbb{P}(Z_n = x_n^n)$.

The result is a union of contiguous rectangles, and, for every $a < b$, $\mathbb{P}(a < Z_n \leq b)$ is the sum of the ares of the rectangles centered on points x_m^n for which $a < x_m^n \leq b$. Let $f_n(x)$ be the function defined the upper boundaries of these rectangles:

$$f_n(x) = \begin{cases} \sqrt{npq}C_n^m p^m q^{n-m} & \text{if } x \in [x_m^n - \frac{1}{2\sqrt{npq}}, x_m^n + \frac{1}{2\sqrt{npq}}),\ 0 \leq m \leq n \\ 0 & \text{otherwise.} \end{cases}$$

The union of all the rectangles is just the region between f_n and x-axis,

$$\int_{x_m^n - \frac{1}{2\sqrt{npq}}}^{x_m^n + \frac{1}{2\sqrt{npq}}} f_n(x)dx = \mathbb{P}(Z_n = x_m^n),\ \text{for each } m,$$

and hence

$$\begin{aligned} \mathbb{P}(a < Z_n < b) &= \sum_{m, a < x_m^n < b} \mathbb{P}(Z_n = x_m^n) \\ &= \sum_{m; a < x_m^n < b} \int_{x_m^n - \frac{1}{2\sqrt{npq}}}^{x_m^n + \frac{1}{2\sqrt{npq}}} f_n(x)dx = \int_a^b f_n(x)dx + R_n, \end{aligned}$$

where $|R_n| \leq \frac{1}{\sqrt{npq}} \sup_x |f_n(x)|$. In particular, $\int_{-\infty}^{\infty} f_n(dx)dx = 1 \sum_0^n \mathbb{P}(Z_n = x_m^n) = 1$. From these relations, it follows that f_n is a probability density whose distribution can be used to approximate the distribution of Z_n. In fact, f_n is the probability density of the random variable

$$Z_n + \frac{U}{\sqrt{npq}}$$

where U is uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$ and is independent of Z_n. This is easy to see, because, given $Z_n = x_m^n$, $U/\sqrt{npq}$ is uniformly distributed on $(x_m^n - \frac{1}{2\sqrt{bpq}}, x_m^n + \frac{1}{2\sqrt{npq}})$, and thus its density is flat on this interval and must integrate over this interval to $\mathbb{P}(Z_n = x_m^n)$. One can think of $Z_n + U/\sqrt{npq}$ as a perturbed, smoothed version of Z_n, that approximates Z_n better and better as n gets larger.

The amazing fact, discovered essentially by *DeMoivre*, is that the density $f_n(x)$ converges to a simple, normal limit.

Theorem 4.1.1 *(DeMoivre-Laplace local limit theorem)* For any $K < +\infty$,

$$\lim_{n\to\infty} \frac{f_n(x)}{e^{-x^2/2}/\sqrt{2\pi}} = 1, \text{ uniformly for } |x| \leq K \tag{4.2}$$

The rest of this section is devoted to a proof of above theorem. By the definition of f_n, the limit statement, (4.2), in this theorem is equivalent to

$$\lim_{n\to\infty} \frac{\sqrt{npq} C_n^m p^m q^{n-m}}{e^{-(x_m^n)^2/2}/\sqrt{2\pi}} = 1, \text{ uniformly for } |x_m^n| \leq K, \text{ for any } K < +\infty, \tag{4.3}$$

and this is what we worked toward proving.

There are two ingredients. The first comes from an application of *Taylor's formula* with remainder:

$$(1+x)\ln(1+x) = x + x^2/2 + R(x), \text{ where } R(x) = O(x^3). \tag{4.4}$$

The second is the following refinement of *Stirling's formula*,

$$n \sim \sqrt{2\pi n} n^n e^{-n} e^{\theta_n}, \text{ where } 0 < \theta_n < \frac{1}{12n}.$$

It follows easily from the definition $x_m^n = \frac{m-np}{\sqrt{npq}}$ that

$$\frac{m}{np} = 1 + x_m^n \sqrt{\frac{q}{np}}, \ \frac{n-m}{nq} = 1 - x_m^n \sqrt{\frac{p}{nq}}.$$

As a consequence, if $|x_m^n|$ is kept below any constant K as $n \to \infty$, then $m \to \infty$ and $n - m \to \infty$ as well. This is important in what follows.

In the next calculation we firstly apply *Stirling's formula*, and then algebraic manipulations:

$$\sqrt{npq}C_n^m p^m q^{n-m} = \sqrt{npq}\frac{n!}{m!(n-m)!}p^m q^{n-m} \tag{4.5}$$

$$= \sqrt{npq}\frac{\sqrt{n}}{\sqrt{2\pi m(n-m)}}\frac{n^n}{m^m(n-m)^{n-m}}p^m q^{n-m}e^{\theta_n-\theta_m-\theta_{n-m}} \tag{4.6}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\frac{m}{np}\frac{n-m}{nq}}}(\frac{m}{np})^{-m}(\frac{n-m}{nq})^{-(n-m)}e^{\theta_n-\theta_m-\theta_{n-m}} \tag{4.7}$$

Clearly, if $|x_m^n|$ is always kept less than some fixed $K < \infty$ as $n \to \infty$,

$$\frac{1}{\sqrt{\frac{m}{np}\frac{n-m}{nq}}} = \frac{1}{1+x_m^n\sqrt{\frac{q}{np}}}\frac{1}{1-x_m^n\sqrt{\frac{p}{nq}}} \to 1 \text{ uniformly as } n \to \infty. \tag{4.8}$$

By the same token,

$$|\theta_n - \theta_n - \theta_{n-m}| \le \frac{1}{12n}\left[\frac{1}{n}+\frac{n}{m}+\frac{n}{n-m}\right]$$

$$= \frac{1}{12n}\left[1+(1+x_m^n\sqrt{\frac{q}{np}})^{-1}+(1-x_m^n\sqrt{\frac{p}{nq}})^{-1}\right].$$

Thus,

$$e^{\theta_n-\theta_m-\theta_{n-m}} \to 1 \text{ uniformly as } n \to \infty, \text{ if } |x_m^n| \le K. \tag{4.9}$$

The heart of the matter is the term:

$$(\frac{m}{np})^{-m}(\frac{n-m}{nq})^{-(n-m)} = \exp\{-np\cdot\frac{m}{np}\ln(\frac{m}{np}) - nq\cdot\frac{n-m}{nq}\ln(\frac{n-m}{nq})\}$$

We now use *Taylor expansion* in the argument of the exponent:

$$-np\cdot\frac{m}{np}\ln(\frac{m}{np}) - nq\cdot\frac{n-m}{nq}\ln(\frac{n-m}{nq}) \tag{4.10}$$

$$=np(1+x_m^n\sqrt{\frac{q}{np}})\ln(1+x_m^n\sqrt{\frac{q}{np}}) - nq(1-x_m^n\sqrt{\frac{p}{nq}})\ln(1-x_m^n\sqrt{\frac{p}{nq}}) \tag{4.11}$$

$$=\frac{(x_m^n)^2}{2} - npR(x_m^n\sqrt{\frac{q}{np}}) - nqR(x_m^n\sqrt{\frac{p}{nq}}). \tag{4.12}$$

Since $R(x) = O(x^3)$, uniformly, as long as $|x_m^n| < K$ for some fixed $K < +\infty$, it follows from (4.10) that,

$$\frac{(\frac{m}{np})^{-m}(\frac{n-m}{nq})^{-(n-m)}}{e^{-(x_m^n)^2/2}} \to 1, \text{ uniformly as } n \to \infty \text{ for } |x_m^n| < K. \tag{4.13}$$

The result of (4.8), (4.9) and (4.13) applied to (4.8) prove (4.3). This completes the proof of the local limit theorem.

4.2 Weak convergence and convergence in distribution

Our ultimate goal is to prove the *Linderberg-Feller CLT*. At the same time, we want to develop it in the framework of convergence in distribution. This section introduces this notion of convergence in a general setting.

Definition 4.2.1 A sequence $\{\mathbb{F}_n\}$ of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to *converge weakly* to the probability measure $\mathbb{F}$ if

$$\int g(x)\mathbb{F}_n(dx) \to \int g(x)\mathbb{F}(dx) \text{ for all bounded, continuous functions } g \text{ on } \mathbb{R}^d. \tag{4.14}$$

In this case we write $\mathbb{F}_n \xrightarrow{W} \mathbb{F}$. If $\{X_n\}$ is a sequence of $\mathbb{R}^d$-valued random variables, we say that $\{X_n\}$ converges in distribution to X, written $X_n \xrightarrow{d} X$, if $\mathbb{F}_{X_n} \xrightarrow{W} \mathbb{F}_X$.

For a random variable X, $\mathbb{E}[g(X)] = \int g(x)\mathbb{F}_X dx$. Thus $X_n \xrightarrow{d} X$ it and only if $\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$ for every bounded, continuous g.

Remark 4.2.1

1. This definition, it is not necessary that the random variables $\{X_n\}$ are defined on the same probability space. The statement is purely at the level of the individual distributions. In particular, the joint distributions among X_1, X_2, ... are irrelevant;

2. Let $(S, \mathcal{B}(S))$ be a metric space together with its *Borel* σ-algebra. An S-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \mapsto S$ which is measurable in the sense that $X^{-1}(U) \in \mathcal{F}$ for all $U \in \mathcal{B}(S)$. The definition of convergence in distribution continues to make sense for S-valued; one needs only to require that (4.14) holds for each bounded continuous $g : S \mapsto \mathbb{R}$. This is not a generalization for its own sake. Stochastic processes induces a probability measure on infinite dimensional spaces of paths, for example a space of continuous paths, with a topology. The definitions here can then be applied to studying convergence in distribution of a sequence of stochastic processes. In this chapter, we deal almost exclusively with convergence in distribution for real-valued random variables.

Example 4.2.2 As usual, if Y is a random variable, F_Y shall denote its cumulative distribution. It will be shown that $X_n \xrightarrow{d} X$ is equivalent to

$$\lim_n F_{X_n}(b) = F_X(b) \text{ for every continuity point } b \text{ of } F_X.$$

A simple argument shows that this in turn is equivalent to

$$\lim_n \left[F_{X_n}(b) - F_X(a)\right] = F_X(b) - F_X(a)$$

for any finite continuity points a and b, $a < b$ of F_X. Therefore the Central Limit Theorem stated above as Theorem 4.0.3 can be rephrased as follows: if X_1, X_2, ... are i.i.d. with finite

mean μ, and finite variance, σ^2, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z,$$

where Z is standard normal.

Example 4.2.3 Recall that if μ and ν are two measures on $(\Omega, \mathcal{F})$, the total distance between them is:

$$\|\mu - \nu\|_{TV} := \sup\left\{\sum_1^n |\mu(A_i) - \nu(A_i)|\right\}.$$

where the supremum is taken over all finite, measurable partition of Ω. An equivalent definition is:

$$\|\mu - \nu\|_{TV} := \sup\left\{|\int f d(\mu - \nu)|; |f| \leq 1\right\}.$$

In the case of two measures defined on $\mathbb{R}^d$ by densities with respect to *Lebesgue measure*, say, $\mu(A) = \int_A f(x)dx$ and $\nu(A) = \int_A g(x)dx$, the total variation norm is simply $\|\mu - \nu\|_{TV} = \int |f(x) - g(x)|dx$, that is, the L^1-distance between the densities. If μ and ν are probability measure which concentrate all their respective probability mass on disjoint sets, then $\|\mu - \nu\|_{TV} = 2$.

Convergence in total variation certainly implies convergence in distribution, because if $\|\mathbb{F}_{X_n} - \mathbb{F}_X\|_{TV} \to 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} \int g(x)dF_{X_n}(x) = \int g(x)dF_X(x)$$

for any bounded, Borel g. But the converse is not true. The crucial difference between convergence in distribution and in total variation comes down to topology; total variation distance ignores topological closeness, but converging in distribution can occur between random variables whose supports converging even if they are supported on disjoint sets.

Point-wise convergence of density functions implies convergence in distribution. This is the content of the new two results.

Theorem 4.2.4 *(Scheffe's theorem)* Let $\{f_n(x)\}$ be a sequence of probability density functions. Assume that $f(x) = \lim_{n\to\infty} f_n(x)$ for *Lebesgue almost-everywhere* x and that f is also a probability density function. Let $\mathbb{F}_n$, $n \geq 1$, and $\mathbb{F}$ denote the distribution measures corresponding to these densities. Then

$$\|\mathbb{F}_n - \mathbb{F}\|_{TV} \to 0 \text{ as } n \to \infty.$$

Corollary 4.2.5 Under the hypothesis of *Scheffe's theorem*, $\mathbb{F}_n \xrightarrow{d} \mathbb{F}$.

Proof. (proof of Scheffe's theorem) Observe that

$$\|\mathbb{F}_n - \mathbb{F}\|_{TV} = \int |f_n(x) - f(x)|dx$$

Let $A_n := \{x : f(x) > f_n(x)\}$. Because f_n and f are probability densities,

$$\int_{A_n^c} f_n(x)dx = 1 - \int_{A_n} f_n(x)dx \text{ and } \int_{A_n^c} f(x)dx = 1 - \int_{A_n} f(x)dx$$

Thus,

$$\begin{aligned}\int |(f - f_n)(x)|dx &= \int_{A_n} (f - f_n)(x)dx + \int_{A_n^c} (f_n - f)(x)dx \\ &= 2\int |(f - f_n)(x)|\mathbf{1}_{A_n}(x)dx.\end{aligned}$$

The integrand of the final integral is dominated by f and tends to 0 almost everywhere as $n \to \infty$. Thus, by dominated convergence theorem,

$$\lim_{n\to\infty} \int |(f - f_n)(x)|dx = 0.$$

□

The first main theorem of this section establishes alternative criteria for weak convergence. It is sometimes referred to as the *portmanteau theorem*.

Theorem 4.2.6 Let $\mathbb{F}_n$, $n \geq 1$, and $\mathbb{F}$ be probability distributions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The following are equivalent:

(a) $\mathbb{F}_n \xrightarrow{W} \mathbb{F}$ *(weakly)*;

(b) $\lim_{n\to\infty} \int g(x)d\mathbb{F}_n(dx) = \int g(x)\mathbb{F}(dx)$ for every uniformly continuous bounded g;

(c) $\liminf \mathbb{F}_n(G) \geq \mathbb{F}(G)$ for every open set G;

(d) $\limsup \mathbb{F}_n(C) \leq \mathbb{F}(C)$ for every closed set C;

In addition, if $d = 1$, and F_n and F denote the cumulative distribution function of $\mathbb{F}_n$ and $\mathbb{F}$, respectively, then statements above are equivalent to:

(e) $F_n(b) \to F(b)$ at every continuity point b of F.

Characterization (e) of this theorem expresses the notion of convergence in distribution in perhaps its most directly understandable and intuitive form. It implies, for instance if $X_n \xrightarrow{d} X$, then, whenever $a < b$ are continuity points of F,

$$\lim_{n\to\infty} \mathbb{P}(a < X_n \leq b) = \lim_{n\to\infty} F_{X_n}(b) - F_{X_n}(a) = F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b). \qquad (4.15)$$

Conversely, if (4.15) holds for all continuity points, $a < b$, of F_X, then (e) holds and $X_n \xrightarrow{d} X$. In light of the equivalence between (4.15) and characterization (e), the *Central Limit Theorem* maybe rephrased as saying that:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z.$$

where Z is a standard, normal random variable, because every $b \in \mathbb{R}$ is a continuity point of the normal distribution.

The implication, (a) to (c), is easy to understand on intuitive grounds. Let G be open, let $\delta > 0$, and suppose $\mathbb{F}_n$ assigns mass at least δ to a set of points A_n tending to the boundary of G. Thus probability mass of at least δ leaves G in the limit. For (d), the situation is opposite: new probability mass can enter a closed set in the limit.

We will prove all equivalence for the case $d = 1$ (probability distribution on $\mathbb{R}$). The equivalence of (a)-(d) on $\mathbb{R}^d$, $d > 1$ is a simple extension of our arguments. Actually, these equivalence all hold even when $\mathbb{R}^d$ is replaced by a *Polish space* – a complete, separable, metric space. This is explained in more details later in this section.

Proof. ((a) implies (b)): This is immediate.

((b) implies (c)): Given any open set G, let $g_k(x) = \min\{1, k\text{dis}(x, G^{\mathcal{C}})\}$ where $\text{dist}(x, G\mathcal{C}) = \inf\{|x - y|; y \in G^{\mathcal{C}}\}$. Each g_k is uniformly continuous and $g_k \uparrow \mathbf{1}_G$ as $k \to \infty$. Thus, using (b),

$$\liminf_{n\to\infty} \mathbb{F}_n(G) \geq \liminf_{n\to\infty} \int g_k(x)\mathbb{F}_n(dx) = \int g_k(x)\mathbb{F}(dx)$$

for every k. Now take limits as $k \to \infty$ and use the dominated convergence theorem to obtain,

$$\liminf_{n\to\infty} \mathbb{F}_n(G) \geq \int \mathbf{1}_G(x)\mathbb{F}(dx) = \mathbb{F}(G).$$

((c) if and only if (d)): Observe that $\mathbb{F}_n(C) = 1 - \mathbb{F}_n(C^{\mathcal{C}})$. Applying (c) to $\mathbb{F}_n(C^{\mathcal{C}})$ gives (d). Likewise (d) implies (c).

((d) implies (e)): For every b, (d) implies

$$\limsup F_n(b) = \limsup \mathbb{F}_n((-\infty, b]) \leq \mathbb{F}((-\infty, b]) = F(b).$$

On the other hand (d) implies (c) and so

$$\liminf F_n(b) \geq \liminf \mathbb{F}_n((-\infty, b)) \geq \mathbb{F}((-\infty, b)).$$

But if b is a continuity point of F, we know $F(b) = \mathbb{F}((-\infty, b))$, and so combining these two statements yields,

$$\limsup F_n(b) \leq F(b) \leq \liminf F_n(b)$$

and (e) follows.

((e) implies (c)): Since any open set in $\mathbb{R}$ is a countable union of disjoint open intervals, it suffices to show that (e) implies that $\liminf \mathbb{F}_n((a, b)) \geq \mathbb{F}((a, b))$ for $-\infty \leq a < b$. Let (a_k, b_k) be a sequence of open intervals such that $(a_k, b_k) \subset (a, b)$ for every k, $a_k \downarrow a$, $b_k \uparrow b$, and a_k and b_k are continuity points of F for all k. Then, using (e),

$$\begin{aligned}\liminf \mathbb{F}_n((a, b)) \geq & \liminf \mathbb{F}_n((a_k, b_k)) \\ = & \liminf F_n(b_k) - F_n(a_k) \\ = & F(b_k) - F(a_k) = \mathbb{F}((a_k, b_k))\end{aligned}$$

for every k. Since $\mathbb{F}((a_k, b_k)) \uparrow \mathbb{F}((a, b))$ as $k \to \infty$, taking $k \to \infty$ gives $\liminf \mathbb{F}_n((a, b)) \geq \mathbb{F}((a, b))$ as desired.

((c) implies (a)): (c) implies (d) and so for any A,

$$\mathbb{F}(\text{int}(A)) \leq \liminf \mathbb{F}_n(A) \leq \limsup \mathbb{F}_n(\bar{A}) \leq \mathbb{F}(\bar{A}).$$

It follows that if $\mathbb{F}(\text{int}(A)) = \mathbb{F}(\bar{A})$, or, equivalently, $\mathbb{F}(\bar{A} - \text{int}(A)) = 0$, then $\lim_{n\to\infty} \mathbb{F}_n(A) = \mathbb{F}(A)$. This is a useful preliminary. Now let g be a bounded continuous function. We claim that for every positive integer k there is a sequence of simple functions $\{g_k\}$ such that $\|g_k - g\| := \sup_x |g_k(x) - g(x)| \leq \frac{1}{k}$ for each k and such that

$$\lim_{n\to\infty} \int g_k(x) \mathbb{F}_n(dx) = \int g_k(x) \mathbb{F}(dx) \text{ for each } x. \tag{4.16}$$

This will suffice to complete the proof because:

$$\begin{aligned}\left| \int g(x) \mathbb{F}_n(dx) - \int g(x) \mathbb{F}(dx) \right| \leq & \int |g_k(x) - g(x)| \mathbb{F}_n(dx) \\ & + \int |g_k(x) - g(x)| \mathbb{F}(dx) + \left| \int g_k(x) \mathbb{F}_n(dx) - \int g_k(x) \mathbb{F}(dx) \right| \\ \leq & \frac{2}{k} + \left| \int g_k(x) \mathbb{F}_n(dx) - \int g_k(x) \mathbb{F}(dx) \right|.\end{aligned}$$

Thus from (4.16),

$$\limsup_{n\to\infty} \left| \int g(x) d\mathbb{F}_n(dx) - \int g(x) \mathbb{F}(dx) \right| \leq \frac{2}{k}.$$

Since k is arbitrary, we can take $k \to \infty$ and recover (a).

It remains to construct the sequence g_k. Without loss of generality, assume $-1 < g(x) < 1$ for all x. The set of values $N = \{y | \mathbb{F}(g^{-1}(\{y\})) > 0\}$ is at most countable since the sets $g^{-1}(\{y\})$ are disjoint for different y. Therefore, no matter what $k > 0$ is, there is a partition $a_0 = -1 < a_1 < \cdots < a_{m_k} = 1$ so that $\mathbb{F}(g^{-1}(\{a_i\})) = 0$ for every i and $a_1 - a_{i-1} < \frac{1}{k}$ for every i. For each i, set $A_i = \{x | a_i < g(x) \leq a_{i+1}\}$ and define $g_k(x) = \sum a_i \mathbf{1}_{A_i}(x)$. It is clear that $\|g_k - g\| \leq \frac{1}{k}$. Hence, we need only to show (4.16). For this, note that $\bar{A}_i - \text{int}(A_i) \subset g^{-1}(\{a_{i+1}\}) \cup g^{-1}(\{a_i\})$ and so $\mathbb{F}(\bar{A}_i - \text{int}(A_i)) = 0$ and hence $\mathbb{F}_n(A_i) - \mathbb{F}(A_i)$ for all k. Thus,

$$\int g_k(x)\mathbb{F}_n(dx) = \sum a_i \mathbb{F}_n(A_i) \to \sum a_i \mathbb{F}(A_i) = \int g_k(x)\mathbb{F}(dx)$$

as $n \to \infty$, proving (4.16). □

Since convergence in distribution of X_n to X is a statement only about the distributions, there is no need to have the random variables defined on the same probability space. But if there are, we also have notions of convergence almost surely, or in probability, or in L^p. These types of convergence all imply convergence in distribution.

Theorem 4.2.7 Convergence in probability implies convergence in distribution. Consequently, almost sure and L^p convergence also imply convergence in distribution.

Proof. Let $X_n \xrightarrow{P} X$. To show convergence in distribution it suffices by (e) of theorem above to show $\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$ for every bounded, uniformly continuous function f. So let g be uniformly continuous, and for ϵ, let $\delta > 0$ be such that $|x - y| < \delta$ implies $|g(x) - g(y)| < \epsilon$. Then

$$\mathbb{E}\Big[|g_n(X) - g(X)|\Big] \leq \epsilon + 2\|g\|_\infty \mathbb{P}\left(|X_n - X| \geq \delta\right),$$

where $\|g\|_\infty = \sup_x |g(x)|$. As $n \to \infty$, the second term goes to 0 and so $\limsup_{n\to\infty} \mathbb{E}\Big[|g(X_n) - g(X)|\Big] \leq \epsilon$. Taking $\epsilon \downarrow 0$ completes the proof. □

There is a converse of sorts.

Theorem 4.2.8 Suppose $X_n \xrightarrow{P} X$. There is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting random variables $X_1', X_2', ...$ and X' such that $F_{X_n'} = F_{X_n}$ for every n, $F_{X'} = F_X$, and $\lim_{n\to\infty} X_n' = X$ almost surely.

Proof. This can be proved using the quantile coupling. Recall that if F is a cumulative distribution function and $F^{-1}(x) = \inf\{y; F(y) \geq x\}$, then if U is uniformly distributed on $(0,1)$, $F^{-1}(U)$ is a random variable whose cumulative distribution function is F. Fix a probability space supporting such a random variable U, and let $X_n' := F_{X_n}^{-1}(U)$ for each n, and $X' = F_X^{-1}(U)$ The proof will be finished if we show X_n' converges to X almost surely (which we leave readers to complete). □

The second main theorem of this section has to do with extracting weakly convergence subsequences. A family of probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be *relatively compact* if any sequence drawn from the family contains a subsequence converging weakly to a probability measure.

Definition 4.2.2 A family $\{\mathbb{F}_\alpha; \alpha \in \mathcal{A}\}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is *tight* if for every $\epsilon > 0$ there exists a compact set K such that $\mathcal{F}_\alpha(K) \geq 1 - \epsilon$ for every $\alpha \in \mathcal{A}$. We say that a collection of random variables or a collection of cumulative distribution functions is tight if the corresponding collection of distribution measures are tight.

Example 4.2.9 Assume that $\sup_n |\mathbb{E}[Y_n]| < \infty$ and $\sup_n \mathrm{Var}(Y_n) < \infty$. Then $\{Y_n\}$ is a tight family. This is a simple consequence of *Chebyshev's inequality*. Let $m = \sup_n |\mathbb{E}[Y_n]|$. Then

$$\sup_n \mathbb{P}\left(|Y_n| > m + M\right) \leq \sup_n \mathbb{P}\left(|Y_n - \mathbb{E}[Y_n]| > M\right) \leq \frac{\sup_n \mathrm{Var}(Y_n)}{M^2},$$

By choosing M large enough, the probability Y_n lies outside of the compact set $[-m-M, m+M]$ can be made arbitrarily small, uniformly over n.

This is precisely the situation we encounter in the basic Central Limit Theorem, Theorem 4.0.3. There

$$Z_n = \frac{\sum_1^n X_i - \mu}{\sigma\sqrt{n}}$$

was defined precisely so that it has mean zero and variance one for each n. Hence, $\{Z_n\}$ is a tight family of random variables.

Theorem 4.2.10 *(Helly-Bray theorem)* Let $\{\mathbb{F}_n\}$ be tight. Then there is a subsequence $\{\mathbb{F}_{n_k}\}$ and a probability measure $\mathbb{F}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mathbb{F}_{n_k} \xrightarrow{W} \mathbb{F}$.

Proof. For each n, let $F_n(x) := \mathbb{F}_n\left((-\infty, x]\right)$ be the cumulative distribution function associated to $\mathbb{F}_n$. Let D be a countable dense subset of $\mathbb{R}$, labeled $D = \{x_1, x_2, ...\}$. We apply a diagonalization procedure. Since $F_n(x_1)$ is bounded sequence, there exists a subsequence $\{n_{1k}; k \geq 1\}$ such that $\lim_{k\to\infty} F_{n_{1k}}(x_1)$ exists. Continuing inductively, for each $j \geq 1$, let $\{n_{j+1,k}; k \geq 1\}$ be a subsequence of $\{n_{j+1,k}; k \geq 1\}$ be a subsequence of $\{n_{jk}; k \geq 1\}$ such that $\lim_{k\to\infty} F_{n_{j+1,k}}(x_{j+1})$ exists. Then for the subsequence $\{n_{kk}; k \geq 1\}$,

$$\widetilde{F}(x) := \lim_{k\to\infty} F_{n_{kk}}(x) \text{ exists for every } x \in D.$$

Now define,

$$F(x) := \lim_{y \downarrow x, y \in D, y > x} \widetilde{F}(y).$$

Clearly $0 \leq F(x) \leq 1$ for all x. We claim that F is non-decreasing, right continuous and that $\lim_{j\to\infty} F_{n_{kk}}(x) = F(x)$ at every continuity point x of F. It is clear that $\widetilde{F}$ is non-decreasing, since F_n is for every n. The right-continuity of F then follows by how it is defined. For any x and $\epsilon > 0$, and $y \in D$ such that $x \leq y$,

$$\limsup_{k\to\infty} F_{n_{kk}}(x) \leq \limsup F_{n_{kk}}(y) = \widetilde{F}(y).$$

and by letting $y \downarrow x$,

$$\limsup_{k\to\infty} F_{n_{kk}}(x) \leq F(x)$$

If $z < y < x$ and $y \in D$, then $F(z) \leq \lim_{k\to\infty} F_{n_{kk}}(y) \leq \liminf_{k\to\infty} F_{n_{kk}}(x)$. Thus if x is a continuity point of F, then by letting $z \uparrow x$,

$$F(x) \leq \liminf_{j\to\infty} F_{n_{kk}}(x)$$

It follows that at a continuity point,

$$F(x) = \lim_{k\to\infty} F_{n_{kk}}(x). \tag{4.17}$$

So far we have not used tightness. But because of tightness, for any $\epsilon > 0$, there are a and b, $a < b$, that are continuity points of F such that $F_n(b) - F_n(a) > 1 = \epsilon$ for all n. Thus, from (4.17), $F(b) - F(a) > 1 - \epsilon$. Since ϵ is arbitrary, $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

We have therefore shown that F is a cumulative distribution function. There is a unique probability distribution $\mathbb{F}$ such that $F(x) = \mathbb{F}((-\infty, x])$. By *portmanteau theorem*, part (e), $\mathbb{F}_{n_{kk}} \xrightarrow{W} \mathbb{F}$. □

The *Helly-Bray theorem* supplies a strategy for proving convergence in distribution of a sequence $\{Y_n\}$. Show first that $\{\mathbb{F}_{Y_n}\}$ is tight. Show next that any subsequence converging in distribution, must converge to the same limit, call it $\mathbb{F}$. It then follows that $\mathbb{F}_{Y_n} \xrightarrow{W} \mathbb{F}$. If it did not, we could find an $\epsilon > 0$, a continuous, bounded g, and a subsequence $\{n_k\}$ such that

$$\left| \int g(d\mathbb{F}_{Y_n} - d\mathbb{F}) > \epsilon \right|$$

for all k. But this is a contradiction, because, by assumption, there is a subsequence of $\{n_k\}$ along which we have convergence in distribution to F.

4.3 Connection to topology and functional analysis

Weak convergence defines a topology on the space of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, namely the smallest topology that makes the map $\mathbb{F} \to \int g(x)\mathbb{F}(dx)$ continuous for every bounded g. In fact, this is a metric topology. We shall not employ this topology explicitly, but it is worthwhile at this point to see how weak convergences relates to dual spaces and dual topologies, as introduced in functional analysis.

Let C_0 denote the set of continuous functions on $\mathbb{R}$ vanishing at infinity and make C_0 a *Banach space* by endowing it with supremum norm. The topological dual, C_0^*, of C_0 is the set of all continuous linear functions on C_0. Now, any bounded signed measure ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defines a linear function on C_0 by

$$\nu(g) := \int g d\nu,$$

and it is a continuous mapping because, $|\int g d\nu| \leq \sup_x |g(x)||\nu(\mathbb{R})|$, where $|\nu|$ is the sum of the positive and negative parts of ν. Thus, bounded signed measures are contained in C_0^*. *Riesz's theorem* states, in fact, that C_0^* equals the space of all bounded, signed *Borel* measures on $\mathbb{R}$. In functional analysis, the *weak-* topology on* C_0^* is defined to be the smallest topology with respect to which the maps $\mu \mapsto \int_{\mathbb{R}} g(x)\mu(dx)$, where $\nu \in C_0^*$, are continuous for every $g \in C_0$. In other words, the sequence of bounded measure μ_n converges to μ in the weak-* topology if and only if

$$\int g(x)\mu_n(x) \to \int g(x)\mu(dx) \text{ for every } g \in C_0.$$

Now the set of probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a subset of C_0^*. And certainly $\mathbb{F}_n \xrightarrow{W} \mathbb{F}$ implies weak-* convergence of $\mathbb{F}_n$ to $\mathbb{F}$ as an element of C_0^*. It turns out that $\mathbb{F}_n \xrightarrow{W} \mathbb{F}$ if and only if $\mathbb{F}_n \to \mathbb{F}$ in the weak-* topology. Thus, what we have called weak convergence is what the functional analyst would weak-* convergence. This is potentially confusing because weak convergence has a different meaning in functional analysis! But the use in probability theory of 'weak convergence' for 'weak-* convergence' has been sanctified by too long and widespread a use to be changed.

Remark 4.3.1 The *Helly-Bray theorem* is actually a corollary of a general result in functional analysis called the *Banach-Alaoglu theorem*, and it is instructive to discuss the connection. Let B be a *Banach space* and let B^* be its topological dual. Then the *Banach-Alaoglu theorem* states that the unit ball in B^* is compact in the weak-* topology. The space of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a subset of the unit ball in the dual, C_0^*. Therefore, any sequence $\{\mathbb{F}_n\}$ of probability measures contains a subsequence converging in the weak-* sense to an element of $\mathbb{F}$ of the unit ball of C_0^*. As an element of C_0^*, $\mathcal{F}$ is a bounded signed measure and it clearly must be non-negative measure. If $\{\mathbb{F}_n\}$ is tight, then, just as in the proof we gave of the *Helly-Bray theorem*, $\mathbb{F}$ must be a probability measure.

The discussion up to this point has dealt only with real-valued random variables, or, more precisely, with probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The restriction to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ was made only for the sake of keeping the exposition more concrete. But the definition of weak convergence has nothing specifically about the structure of $\mathbb{R}$, beyond a topology which provides a set of continuous functions. Hence we could frame the definition of weak convergence in the general context of probability measures on topological spaces. In practice, we restrict ourselves to metric spaces.

Definition 4.3.1 Let μ_n, $n \geq 1$ and μ be probability measures on the *Borel sets* of a metric space $\mathcal{S}$. We say that μ_n converges weakly to μ if $\int g(x)\mu_n(dx) \to \int g(x)\mu(dx)$ for every bounded continuous function g on $\mathcal{S}$.

The portmanteau theorem goes through pretty much in its entirety. We need only make an adjustment for statement (e) of the theorem, because for a probability measure on a general metric space there is not a concept of cumulative distribution function. To replace (e), we make the following definition. A Borel subset A of $\mathcal{S}$ is said to be μ-continuity set if $\mu(\bar{A} - \text{int}(A)) = 0$.

Theorem 4.3.2 Let μ_n, $n \geq 1$ be probability distributions on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, where $\mathcal{S}$ is a metric space. Let

(a) $\mu_n \xrightarrow{W} \mu$;

(b) $\lim_{n\to\infty} \int g(x)\mu_n(dx) = \int g(x)\mu(dx)$ for every uniformly continuous, bounded g;

(c) $\liminf \mu_n(G) \geq \mu(G)$ for every open set G;

(d) $\limsup \mu_n(C) \leq \mu(C)$ for every closed set C;

(e) $\lim \mu_n(A) = \mu(A)$ for every μ-continuity set.

The definition of *tightness* can also be generalized to the case of a metric space $\mathcal{S}$ and again serves to characterize relative compactness.

Definition 4.3.2 A single probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ is tight if for every $\epsilon > 0$ there is a compact set K such that $\mu(K) > 1 - \epsilon$. A family $\mathcal{M}$ of probability measures on $\mathcal{S}$ is uniformly tight if for every ϵ, there exists a compact K such that $\mu(K) > 1 - \epsilon$ for all $\mu \in \mathcal{M}$.

One might think that tightness of a probability measure might be an atypical property if $\mathcal{S}$ is, say, an infinite dimensional *Hilbert* or *Banach* space. Surprisingly, this is not the case.

Theorem 4.3.3 If $\mathcal{S}$ is a complete, separable metric space, then every probability measure on $\mathcal{S}$ is compact.

The connection of uniform tightness to relative sequential compactness work as in the case of $\mathcal{S} = \mathbb{R}$, with some minor refinements.

Theorem 4.3.4 *(Prohorov's theorem)*

(i) If $\mathcal{M}$ is uniformly tight family of probability measures on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ then $\mathcal{M}$ is relatively sequentially compact;

(ii) If $\mathcal{S}$ is complete and separable, then a relatively sequentially compact family of probability measure is tight.

4.4 Characteristic Functions and Convergence in Distribution

4.4.1 Definition and basic properties

The *characteristic function* of a random variable of X is

$$\phi_X(\lambda) := \mathbb{E}[e^{i\lambda X}], \quad \lambda \in \mathbb{R}$$

If $\mathbb{F}_X$ denotes the distribution measure of X,

$$\phi_X(\lambda) = \int e^{i\lambda x} \mathbb{F}(dx),$$

and when X is continuous random variable with density f_x, so that $\mathbb{F}_X(dx) = f_X(x)dx$,

$$\phi_X(\lambda) = \int e^{i\lambda x} f_X(x)dx$$

is the *Fourier transform* of f_X. So more generally, one may think of $\phi_X(\lambda)$ as the *Fourier transform* of the measure $\mathbb{F}_X$. Some elementary properties of characteristic functions are:

(i) $\phi_X(0) = 1$ and $|\phi_X(\lambda)| \leq 1$ for all λ;

(ii) For any random variable $\phi_X(\lambda)$ is a continuous function of λ;

(iii) If $n \geq 0$ is a positive integer and $\mathbb{E}[|X|^n] < +\infty$, then $\phi_X(t)$ is n-times continuously differentiable and the k-th derivative is

$$\frac{d^k}{d\lambda^k}\phi_X(\lambda) = i^k \mathbb{E}[X^k e^{i\lambda X}].$$

One of the most important characteristic function is that of normal random variable. If Z is standard normal random variable, then

$$\phi_Z(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\lambda x} e^{-x^2/2} dx = e^{-\lambda^2/2}. \tag{4.18}$$

This formula says, in effect, that $e^{-x^2/2}$ is an *eigenfunction* of the *Fourier transform*, which is an indication of the importance of normal distribution. Formula (4.18) maybe derived by showing, with an integration by parts, that $\phi_Z'(\lambda) = i\mathbb{E}[e^{i\lambda Z}] = -\lambda\phi_Z(\lambda)$. An integration of differential equation, using $\phi_Z(0) = 1$, yields the formula. When Z is standard normal, $X = \mu + \sigma Z$ is normal with mean 0 and variance σ^2, and thus we get (4.18) that

$$\phi_X(\lambda) = e^{-i\mu\lambda - \lambda^2\sigma^2/2}. \tag{4.19}$$

Another important example is the *Poisson random variable*. If Y has the *Poisson distribution* with mean λ,

$$\phi_Y(\lambda) = \sum_{k=0}^{\infty} e^{i\lambda k} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda(e^{i\lambda}-1)}. \tag{4.20}$$

We noted above that $|\phi_X(\lambda)| \leq 1$ always. If there is a non-zero λ_0 such that $|\phi_X(\lambda_0)| = 1$, then X has special structure. We may assume in this case that $\lambda_0 > 0$, since $\phi_X(-\lambda_0)$ is the complex conjugate of $\phi_X(\lambda_0)$ and hence $|\phi_X(-\lambda_0)| = |\phi_X(\lambda_0)| = 1$. Now, there is real number a such that $\mathbb{E}[e^{i\lambda_0 X}] = e^{i\lambda_0 a}$. Hence,

$$1 = \mathbb{E}[e^{i\lambda_0(X-a)}] = \mathbb{E}[\cos(\lambda_0(X-a))] + i\mathbb{E}[\sin\lambda_0(X-a)].$$

Since 1 is an upper bound to $\cos(x)$, this can only occur if $\mathbb{P}\left(\cos(\lambda_0(X-a)) = 1\right) = 1$. Thus

$$X \text{ takes values in the set } \{a + \tfrac{2\pi n}{\lambda_0}; n \in \mathbb{Z}\} \text{ with probability one.}$$

A random variable with this property is called a *lattice* random variable. In this case, it follows easily that $\phi_X(m\lambda_0) = e^{i\lambda_0 a}$ for all integer m. If $|\phi_X(\lambda_n)| = 1$ along some sequence of non-zero λ_n for which $\lambda_n \to 0$, it would follow that $|\phi_X(\lambda)| = 1$ on a dense set of λ and so, by continuity, that $\phi_X(\lambda) = 1$ for all λ. Suppose this is not the case and let

$$\lambda := \inf\{\lambda > 0; |\phi_X(\lambda)|\} = 1.$$

Then X take values in $\{a + nh; n \in \mathbb{Z}\}$ for some a and for $h = \frac{2\pi}{\lambda}$, and this will be true for no larger values of h. h is called the maximum span of X. If, for some b, $\mathbb{P}(X = b) = 1$, then $|\phi_X(\lambda)| = |e^{i\lambda b}| = 1$ for all λ, and this will be the only case in which $|\phi_X(\lambda)| = 1$. We have proved the following result:

Theorem 4.4.1 If there exists non-zero λ such that $|\phi_X(\lambda)| = 1$. Then either X is a degenerate random variable and $|\phi_X(\lambda)| = 1$ or X is a *lattice* random variable. If h is its maximum span, then $|\phi_X(\lambda)| < 1$ for $0 < \lambda < 2\pi/h$.

4.4.2 Independence and convolution

Let X and Y be independent random variables. Then

$$\phi_{X+Y}(\lambda) = \mathbb{E}[e^{i\lambda(X+Y)}] = \mathbb{E}[e^{i\lambda Y}]\mathbb{E}[e^{i\lambda Y}] = \phi_x(\lambda)\phi_Y(\lambda). \tag{4.21}$$

This perhaps the most important property of characteristic functions in the study of central limit theorem. Of course it generalizes to any finite sum of random variables.

A basic property of the *Fourier transform* is that the *Fourier transform* of the convolution of two functions is the product of their *Fourier transforms*. This is essentially what (4.21) says, but in the context of probability measures. Given two probability measures $\mathbb{F}_1$ and $\mathbb{F}_2$ on the *Borel sets* of $\mathbb{R}$, define their convolution as

$$\mathbb{F}_1 * \mathbb{F}_2(A) = \int_{\mathbb{R}} \mathbb{F}_1(A - x)\mathbb{F}_2(dx).$$

It is easy to check directly that $\mathbb{F}_1 * \mathbb{F}_2$ is a probability measure and that convolution is commutative and associative.

If X_1 and X_2 are independent random variables, then, by *Fubini's theorem*,

$$\begin{aligned}\mathbb{P}\left(X_1 + X_2 \in A\right) &= \int\int \mathbf{1}_A(x_1 + x_2)\mathbb{F}_{X_1}(dx_1)\mathbb{F}_{X_2}(dx_2) \\ &= \int \mathbb{F}_1(A - x_2)\mathbb{F}_{X_2}(dx_2) = \left[\mathbb{F}_{X_1} * \mathbb{F}_{X_2}\right](A).\end{aligned}$$

This generalizes by induction to any finite random variables. If $X_1, X_2, ..., X_n$ are independent, then

$$\mathbb{F}_{X_1+\cdots+X_n} = \mathbb{F}_{X_1} * \cdots * \mathbb{F}_{X_n}.$$

We have interpreted the characteristic function of a random variable as the *Fourier transform* of its distribution measure, and so when $X_1, ..., X_n$ are independent $\phi_{X_1+\cdots+X_n} = \phi_{X_1} \cdots \phi_{X_n}$ says that the *Fourier transform* of a convolution of measures is the product of the *Fourier transforms*. Of course one could derive this by direct calculation.

A particular interesting and useful case to consider is

$$X + \epsilon Z$$

where $\epsilon > 0$ and Z is a standard normal random variable. We shall denote the density of Z by $\rho(z) = (\sqrt{2\pi})^{-1/2}e^{-x^2/2}$. Then the density of ϵZ is $\epsilon^{-1}\rho(z/\epsilon)$ and it shall be denoted by $\rho_\epsilon(z)$. Then,

$$\begin{aligned}\mathbb{F}_{X+\epsilon Z}(A) &= \int\int \mathbf{1}_A(x + z)\rho_\epsilon(z)\mathbb{F}_X(dx) \\ &= \int\int \mathbf{1}_A(z)\rho_\epsilon(z - x)dz\mathbb{F}_X(dx) = \int \mathbf{1}_A(z)\Big[\int \rho_\epsilon(z - x)\mathbb{F}_X(dx)\Big]dz.\end{aligned}$$

It follows immediately that $X + \epsilon Z$ is a continuous random variable with density

$$f_{X+\epsilon Z}(x) = \int \rho_\epsilon(x-y)\mathbb{F}_X(dy). \tag{4.22}$$

(We substituted x for z in the previous equation.) When X itself has a density f_X,

$$f_{X+\epsilon Z} = \int \rho_\epsilon(x-y)f_X(y)dy = \rho_\epsilon * f_X(x)$$

where $g * f$ denotes the usual convolution of the functions.

The family $\{\rho_\epsilon; \epsilon > 0\}$ constitutes what is known as *mollifier*. It can be shown that if $1 \leq p < +\infty$ and if g is a function in $L^p(\mathbb{R})$, then

$$f_\epsilon * g \in C^\infty, \ \forall \epsilon > 0 \text{ and } \lim_{\epsilon \downarrow 0} \|f_\epsilon * g - g\|_{L^p} = 0.$$

It follows immediately that

$$\int |\rho_\epsilon * f - f_X| dx \to 0 \text{ as } \epsilon \downarrow 0.$$

and hence $\|\mathbb{F}_{X+\epsilon Z} - \mathbb{F}_X\|_{TV} \to 0$ as $\epsilon \downarrow 0$, if X is a continuous random variable. If X does not have a density, then $f_{X+\epsilon Z}$ will not converge to a density function, but it still is true that $f_{X+\epsilon Z}$ is infinitely differentiable, and

$$\mathbb{F}_{X+\epsilon Z} \to \mathbb{F} \text{ (weakly) as } \epsilon \downarrow 0.$$

This is a direct consequence of the fact that $\lim_{\epsilon \downarrow 0} X + \epsilon Z = X$ everywhere, and results of convergence implications.

4.5 Final Preparations for CLT

4.6 Fourier Inversion

The characteristic function of $X + \epsilon Z$ is

$$\phi_{X+\epsilon Z}(\lambda) = \phi_{\epsilon Z}(\lambda)\phi_X(\lambda) = e^{-\lambda^2\epsilon^2/2}\phi_X(\lambda).$$

Thus, the addition of the smoothing factor ϵZ also has a dramatic regularization effect on the characteristic function, at least in terms of its integrability properties, since $e^{-\lambda^2\epsilon^2/2}$ decays so rapidly as $\epsilon \downarrow 0$. This is key to the next theorem.

Theorem 4.6.1 Let X be any random variable. Let $a < b$ be continuity points of F_X. Then

$$F_X(b) - F_X(a) = \lim_{\epsilon \downarrow 0} \frac{1}{2\pi} \int e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\frac{e^{-i\lambda b} - e^{-i\lambda a}}{-i\lambda} d\lambda \tag{4.23}$$

As an immediate corollary, the characteristic function is appropriately named – it does truly characterize the distribution of a random variable!

Corollary 4.6.2 If $\phi_X = \phi_Y$, then $F_X = F_Y$.

Proof. (proof of theorem 4.6.1) By applying formula (4.18) for the characteristic function of the standard normal, (switch the roles of λ and x and make a change of variables),

$$\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\lambda^2\epsilon^2/2}d\lambda = \frac{1}{\sqrt{2\pi\epsilon^2}}e^{-x^2/2\epsilon^2}. \tag{4.24}$$

This is the *Fourier inversion formula* for normal densities.

For any fixed x, $e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}e^{i\lambda y}$ is integrable on $\mathbb{R}^2$ as a function of (λ, y) with respect to the measure $m \times \mathbb{F}_X$, where m denotes *Lebesgue measure*. Indeed,

$$\begin{aligned}\int\int |e^{-i\lambda x}e^{-\epsilon^2\lambda^2}e^{i\lambda y}|\mathbb{F}_X(dy)d\lambda &\le \int\int e^{-\epsilon^2\lambda^2/2}\mathbb{F}_X(dy)d\lambda \\ &= \int e^{-\epsilon^2\lambda^2/2}d\lambda < +\infty\end{aligned}$$

Fubini's theorem therefore applies:

$$\begin{aligned}&\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)d\lambda \\ &\qquad = \frac{1}{2\pi}\int \big[\int e^{-i\lambda(x-y)}e^{-\epsilon^2\lambda^2/2}d\lambda\big]\mathbb{F}_X(dy)\end{aligned}$$

Using the inversion formula (4.24) in the inner integral of the last expression and then formula (4.22) for the density of $X + \epsilon Z$ gives

$$\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)d\lambda = \int \frac{e^{-(y-x)^2/2\epsilon^2}/2\epsilon^2}{\sqrt{2\pi\epsilon^2}}d\mathbb{F}_X(y) = f_{X+\epsilon Z}(x) \tag{4.25}$$

Thus, with another application of *Fubini's theorem*,

$$\begin{aligned}F_{X+\epsilon Z}(b) - F_{X+\epsilon Z}(a) &= \int_a^b \frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)d\lambda dx \\ &= \frac{1}{2\pi}\int e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\frac{e^{-i\lambda b} - e^{-i\lambda a}}{-i\lambda}d\lambda.\end{aligned}$$

Because $X + \epsilon Z$ converges everywhere to X and hence in distribution, the *Fourier inversion formula* (4.23) follows by letting $\epsilon \downarrow 0$. □

A *Fourier inversion formula* for densities can be derived as a corollary:

Corollary 4.6.3 If X is a continuous random variable, then

$$\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)d\lambda \text{ converges in } L^1 \text{ to } f_X(x) \text{ as } \epsilon \downarrow 0.$$

If $\phi_X(\lambda)$ is integrable over $\mathbb{R}$. Then X is a continuous random variable and

$$f_X(x) = \frac{1}{2\pi}\int e^{-i\lambda x}\phi_X(\lambda)d\lambda$$

Finally there is also a *Fourier inversion formula* for lattice random variables. It is essentially the formula for the *Fourier coefficients* of a function defined on a finite interval.

Theorem 4.6.4 Let X be a lattice random variable taking values in $\{a+nh; n \in \mathbb{Z}\}$. Let $p_n = \mathbb{P}(X = a+nh)$, $n \in \mathbb{Z}$. Then

$$\phi_X(\lambda) = \sum_{n\in\mathbb{Z}} p_n e^{i\lambda(a+nh)}$$

and

$$p_n = \frac{h}{2\pi}\int_{-\frac{\pi}{h}}^{\pi/h} e^{-i(a+nh)}\phi_X(\lambda)d\lambda. \tag{4.26}$$

4.6.1 Convergence in Distribution

Since $e^{i\lambda x}$ is a bounded, continuous function of x for all λ, if $X_n \xrightarrow{d} X$, then $\phi_{X_n}(\lambda) \to \phi_X(\lambda)$ for all λ. This section is concerned with the opposite implication, proving convergence in distribution from convergence of characteristic functions. This will provide the tool for proving central limit theorem.

Theorem 4.6.5 If $\lim_{n\to\infty}\phi_{X_n}(\lambda) = \phi_X(\lambda)$, then $X_n \xrightarrow{d} X$ (equivalently, $\mathbb{F}_{X_n} \xrightarrow{W} \mathbb{F}_X$).

There is a refinement of this theorem called *Lévy continuity theorem.*

Theorem 4.6.6 Let $\phi(\lambda) = \lim_{n\to\infty}\phi_{X_n}(\lambda)$ exist for all λ and suppose that $\phi(\lambda)$ is continuous at $\lambda = 0$. Then, there is a probability distribution measure $\mathbb{F}$ such that $\phi(\lambda) = \int e^{i\lambda x}\mathbb{F}(dx)$ and $\mathbb{F}_{X_n} \xrightarrow{W} \mathbb{F}$.

Proof. (proof of theorem 4.6.5) Let $\epsilon > 0$. By equation (4.25),

$$f_{X_n+\epsilon Z}(x) = \frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_{X_n}(\lambda)d\lambda.$$

Characteristic functions are bounded uniformly in absolute value by 1, and so by the *dominated convergence theorem*,

$$\begin{aligned}\lim_{n\to\infty} f_{X_n+\epsilon Z}(x) &= \lim_{n\to\infty} \frac{1}{2\pi}\int e^{-i\lambda x} e^{-\epsilon^2\lambda^2/2}\phi_{X_n}(\lambda)d\lambda \\ &= \frac{1}{2\pi}\int e^{-i\lambda x} e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)d\lambda = f_{x+\epsilon Z}(x).\end{aligned}$$

It follows from *Scheffe's theorem* that

$$X_n + \epsilon Z \xrightarrow{d} X + \epsilon Z \text{ for every } \epsilon > 0.$$

We will conclude from this that $X_n \xrightarrow{d} X$. Let f be any bounded, uniformly continuous function and let $\|f\| := \sup_x |f(x)|$. If $\eta > 0$, choose $\delta > 0$ so that $|f(x) - f(y)| < \eta$ if $|x - y| < \delta$. Then,

$$|\mathbb{E}[f(X+\epsilon Z) - f(X)]| \le \eta + \mathbb{E}[|f(X+\epsilon Z) - f(X)|\mathbf{1}_{\{\epsilon|Z|\ge\delta\}}] \le \eta + 2\|f\|\mathbb{P}(|Z| \ge \delta/\epsilon).$$

By the same calculation, $|\mathbb{E}[f(X_n + \epsilon Z) - f(X_n)]| \le \eta + 2\|f\|\mathbb{P}(|Z| \ge \delta/\epsilon)$. Therefore,

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \le 2\eta + 4\|f\|\mathbb{P}(|Z| \ge \delta/\epsilon) + |\mathbb{E}[f(X_n + \epsilon Z)] - \mathbb{E}[f(X + \epsilon Z)]|.$$

But as we have proved that $X_n + \epsilon Z \xrightarrow{d} X + \epsilon Z$,

$$\limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \le 2\eta + 4\|f\|\mathbb{P}(|Z| \ge \delta/\epsilon).$$

Next take $\epsilon \downarrow 0$, and notice that $\mathbb{P}(|Z| \ge \delta/\epsilon) \to 0$; finally, take $\eta \downarrow 0$. It follows that $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$, and this is true for any uniformly continuous, bounded f. Thus, $X_n \xrightarrow{d} X$. □

4.7 Central Limit Theorem

We now have the tools necessary to prove central limit theorems. The basic idea is most clearly revealed in the simplest case of i.i.d. random variables $X_1, X_2, ...$ having finite mean μ and variance σ^2. The central limit theorem for this case was stated previously, and it concerned the limit in the distribution of the sequence,

$$Z_n := \frac{(S_n - n\mu)}{\sigma\sqrt{n}}, \quad n \ge 1,$$

where $S_n = \sum_1^n X_i$.

Let $\phi := \phi_{(X_i-\mu)/\sigma}$. It is the same for all i, because $X_1, X_2, ...$ are identically distributed. Since $(X_i - \mu)/\sigma$ has mean 0 and variance 1, ϕ is twice continuously differentiable and

$$\phi'(0) = \mathbb{E}[(X_i - \mu)/\sigma] = 0 \text{ and } \phi''(0) = -\mathbb{E}[(X_i - \mu)^2/\sigma^2] = 1.$$

Therefore, by *Taylor's expansion*,

$$\phi(\lambda) = 1 - \frac{1}{2}\lambda^2 + R(\lambda)$$

where $R(\lambda) = o(\lambda^2)$ as $\lambda \to 0$.

We will now compute the characteristic function of Z_n in terms of $\phi := \phi_{(X_i-\mu)/\sigma}$. Because $X_1, X_2, ...$ are independent,

$$\phi_{Z_n}(\lambda) = \mathbb{E}\left[\exp i\frac{\lambda}{\sqrt{n}}\sum_1^n \frac{X_i - \mu}{\sigma}\right] = \phi^n(\frac{\lambda}{\sqrt{n}}) = \left(1 - \frac{\lambda^2}{2n} + R(\lambda)\right)^n.$$

Thus,

$$\lim_{n\to\infty}\left(1 - \frac{\lambda^2}{2n} + R(\lambda)\right)^n = e^{-\lambda^2/2.}$$

This is the characteristic function of a standard, normal random variable Z. It follows at once from theorem 4.6.5, or equivalently,

$$\lim_{n\to\infty}\mathbb{P}(a < Z_n \le b) = \int_a^b e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \text{ for all } a < b.$$

This proves the basic *Central Limit Theorem*.

This result inspires a natural question. What can be said if $X_1, X_2, ...$ are independent but not identically distributed? In this case,

$$Z_n = \frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sum_{i=1}^n \mathrm{Var}(X_i)}$$

has mean 0 and variance 1. When does Z_n converge in distribution to a standard normal random variable? Of course, some conditions must be imposed, because by taking $X_2, X_3, ...$ to all be identically zero, $Z_n = (X_1 - \mu_1)/\sigma_1$, for all n and any random variable with mean 0 and variance 1 could be the limit. The right hypothesis is identified in a theorem due to *Lindberg* and *Feller*.

Theorem 4.7.1 Let $X_1, X_2, ...$ be independent random variables with finite means $\mu_i = \mathbb{E}[X_i]$ and variance $\sigma_i^2 = \mathrm{Var}(X_i)$, $i \ge 1$. Let $B_n^2 = \sum_1^n \sigma_i^2$, and set

$$Z_n = \frac{\sum_1^n X_i - \mu_i}{B_n}, \quad n \ge 1.$$

If

$$\lim_{n\to\infty}\frac{1}{B_n^2}\sum_{k=1}^n \mathbb{E}\left[(X_k - \mu_k)^2 \mathbf{1}_{\{|X_k-\mu_k|>\epsilon B_n\}}\right] = 0, \quad \forall \epsilon > 0, \tag{4.27}$$

then,

$$\lim_{n\to\infty} \mathbb{P}(a < Z_n \leq b) = \int_a^b (2\pi)^{-\frac{1}{2}} e^{-z^2/2} dz. \tag{4.28}$$

Conversely, if (4.28) holds, and if

$$\lim_{n\to\infty} \max_{1\leq j\leq n} \mathbb{P}\left(|X_j - \mu_j| \geq B_n\epsilon\right) = 0 \quad \text{for positive } \epsilon. \tag{4.29}$$

then (4.27) must be true.

Condition (4.27) is called *Lindberg's condition*. The converse statement, that *Linderberg's conditions* must hold if the central limit property and (4.29) are true, is due to *Feller*. The next lemma will help give a feeling for what *Linderberg's condition* means.

Lemma 4.7.2 If condition (4.27) holds,

$$\max_{1\leq k\leq n} \frac{\sigma_k^2}{B_n^2} \to 0 \quad \text{as } n \to \infty. \tag{4.30}$$

The rest of this section is devoted to a proof that *Lindeberg's condition* implies the convergence in distribution of Z_n to a standard normal random variable. The converse statement due to *Feller* will not be proved. Clearly, by replacing X_i by $X_i - \mu_i$, there is no loss of generality in assuming that $\mu_i = 0$ for all i, and we shall impose this condition for the rest of the section.

The proof will again come down to showing that $\lim_{n\to\infty} \phi_{Z_n}(\lambda) = e^{-\lambda^2/2}$. The argument, which is technical, requires keeping an accurate track of the error in the *Taylor expansion* of the characteristic function about $\lambda = 0$. This is the purpose of the next lemma.

Lemma 4.7.3 Let X be a mean 0 random variable with finite variance σ^2. For any $\delta > 0$, its characteristic function $\phi_X(\lambda)$ satisfies:

$$|\phi_X(\lambda) - (1 - \frac{\lambda^2}{2}\sigma^2)| \leq \lambda^2 \mathbb{E}\big[X^2 \mathbf{1}_{\{|X|>\delta\}}\big] + \frac{\lambda^3\delta}{6} \mathbb{E}\big[X^2 \mathbf{1}_{\{|X|\leq\delta\}}\big]. \tag{4.31}$$

This lemma says that the variance of each term in the sum $\sum_{i=1}^n X_i - \mu_i$ is vanishingly small compared to the variance of the entire sum, in the limit as $n \to \infty$. For this reason, the *Linderberg-Feller* theorem maybe rephrased roughly as follows: the sum of a large number of independent random variables, none of which dominates the sum, is, when scaled by its variance, approximately normal.

We also need a purely analytical lemma:

Lemma 4.7.4 Let $\{s_{kn}\}$, $1 \leq k \leq n$, $n \geq 1$ be complex numbers satisfying:

$$\lim_{n\to\infty} \max_{1\leq k\leq n} |s_{kn}| = 0, \quad \sup_{n\geq 1} \sum_{k=1}^{n} |s_{kn}| < +\infty, \quad \lim_{n\to\infty} \sum_{k=1}^{n} s_{kn} = \alpha. \tag{4.32}$$

Then,

$$\prod_{k=1}^{n} (1 + s_{kn}) \to e^{\alpha} \text{ as } n \to \infty. \tag{4.33}$$

Now we are ready to prove the *Linderberg's condition*,(4.27), implies that $\phi_{Z_n}(\lambda) \to e^{-\lambda^2/2}$ as $n \to \infty$, proving the convergence in distribution of Z_n to a standard normal.

Since the X_k are independent, the characteristic function of Z_n is

$$\phi_{Z_n}(\lambda) = \prod_{j=1}^{n} \phi_{X_k}(\frac{\lambda}{B_n}).$$

Define the remainder terms $R_k(\lambda) = \phi_{X_n}(\lambda) - (1 - (\lambda^2/2)\sigma_k^2)$. With this notation,

$$\phi_{Z_n}(\lambda) = \prod_{k=1}^{n} \left(1 - \frac{\sigma_k^2 \lambda^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})\right).$$

We shall prove,

$$\lim_{n\to\infty} \sum_{k=1}^{n} |R_k(\frac{\lambda}{B_n})| = 0. \tag{4.34}$$

This will suffice to complete the proof, as we now show. Indeed, (4.34) implies

$$\sup_{n\geq 1} \sum_{k=1}^{n} |-\frac{\sigma_k^2 \lambda^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})| \leq \frac{\lambda^2}{2} + \sup_{n\geq 1} \sum_{k=1}^{n} |R_k(\frac{\lambda}{B_n})| < +\infty,$$
$$\lim_{n\to\infty} \left(\sum_{k=1}^{n} -\frac{\sigma_k^2 \lambda^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})\right) = -\frac{\lambda^2}{2} + \lim_{n\to\infty} \sum_{k=1}^{n} R_k(\frac{\lambda}{B_n}) = -\frac{\lambda^2}{2}.$$

And, because of lemma 4.7.2, equation (4.34) also implies

$$\max_{1\leq k\leq n} |-\frac{\sigma_k^2 \lambda^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})| \leq \lambda^2 \max_{1\leq k\leq n} \frac{\sigma_k^2}{2B_n^2} + \sum_{k=1}^{n} |R_k(\frac{\lambda}{B_n})| \to 0 \text{ as } n \to \infty.$$

We have thus shown that the hypothesis of lemma 4.7.4 are satisfied in the case when $s_{kn} = -\frac{\sigma_k^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})$ and $\alpha = -\lambda^2/2$. By applying this lemma we find that $\lim_{n\to\infty} \phi_{Z_n}(\lambda) = e^{-\lambda^2/2}$, as we claimed.

It remains to show (4.34). For this, apply lemma 4.7.3 with $\delta = \epsilon B_n$. Then, for every $\epsilon > 0$,

$$\sum_{k=1}^{n} |R_k(\frac{\lambda}{B_n})| \leq \frac{\lambda^2}{B_n^2} \sum_{k=1}^{n} \mathbb{E}\big[|X_k|^2 \mathbf{1}_{\{|X_k| > \epsilon B_n\}}\big] + \frac{|\lambda|^3}{6}\epsilon.$$

The *Lindeberg condition* says that as $n \to \infty$, the coefficient of λ^2 in the last expression tends to 0. Hence $\lim_{n\to\infty} \sum_{k=1}^{n} |R_k(\frac{\lambda}{B_n})| < |\lambda^3|\epsilon/2$. Taking $\epsilon \downarrow 0$ proves (4.34). This concludes the proof.