Stockholms universitet

The Ergodic Theorem: A Probabilistic Formulation

Ahmed Bechlaoui

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2025:23
Matematisk statistik
September 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:23**
http://www.math.su.se

The Ergodic Theorem: A Probabilistic Formulation

Ahmed Bechlaoui*

September 2025

Abstract

The main aim of this thesis is to present the probabilistic version of Birkhoff's Ergodic Theorem, whose derivation is carried out from the purely measure theoretic version. A crucial point is to present a rigorous treatment of measure theoretic probability theory, by starting with a more general setting. As an application, it is shown that the Strong Law of Large Numbers is derived as a corollary of the theorem.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ahmed98na14@gmail.com. Supervisor: Daniel Ahlberg, Maria Deijfen.

Acknowledgments

This bachelor's thesis in Mathematical Statistics (15 hp) was carried out at Stockholm University. I would like to thank my supervisors, Daniel Ahlberg and Maria Deijfen, for their support and inspiration during the writing of the thesis. Some AI tools like ChatGPT have been used to write LaTeX code more efficiently, for example when setting up the align environment and refining the code.

Contents

1 Introduction

In elementary probability theory, fundamental results such as the law of large numbers are usually formulated in terms of sequences of random variables or "random numbers". While such statements are intuitive and easily applied, their rigorous foundation often requires a more abstract formulation using measure theory.

Using general measure theory as a starting point, we want to translate the theory in a constructive manner to end up with the measure theoretic language for probability theory, highlighting how random variables, expectations, and convergence theorems can be understood within this general structure. This will help us to understand how the probabilistic version of Birkhoff's Ergodic Theorem follows from the general version, since essentially it is just a translation between mathematical languages. Finally, we demonstrate how the classical Strong Law of Large Numbers arises as a corollary of this theorem.

2 Preliminaries

In this section, we begin by introducing the general construction of measure spaces, which will later enable us to rigorously define probability spaces. Our main focus is on spaces of infinite sequences, for which we aim to construct appropriate event spaces and develop a well-defined notion of probability.

We then provide a brief review of integration theory, which serves as a foundation for formulating Birkhoff's Ergodic Theorem and defining expectations and related concepts. Only the theoretical tools relevant to the main results will be presented.

2.1 Construction of Measure Spaces

Given a set Ω, which is usually referred to as the sample space in the probabilistic language, we define a σ-algebra, a special collection of subsets of the sample space and also know as the event space.

Definition 2.1 (σ-algebra)**.** A *σ-algebra* over a set Ω is a set $\mathcal{A} \subset 2^{\Omega}$ such that:

(i) $\Omega \in \mathcal{A}$,

(ii) if $A \in \mathcal{A}$ then $\Omega \setminus A \in \mathcal{A}$,

(iii) if $A_i \in \mathcal{A}$ for $i \in \mathbb{N}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Given a set Ω and a σ-algebra $\mathcal{A}$ over Ω, the tuple $(\Omega, \mathcal{A})$ is called a *measurable space*. The elements of $\mathcal{A}$ are said to be *measurable sets* (with respect to the given measurable space). This structure allows us to define a measure on $\mathcal{A}$, which is a special type of function that assigns a real number to each measurable set. For instance, in probability theory, we assign probabilities to events using a probability measure.

Example 2.2. The trivial σ-algebra on a set Ω is $\{\emptyset, \Omega\}$.

Example 2.3. Given a set Ω and a collection $\mathcal{C} \subset 2^{\Omega}$, we can easily construct a σ-algebra that contains $\mathcal{C}$ by:

$$\sigma(\mathcal{C}) = \bigcap \left\{\mathcal{A} \subset 2^{\Omega} \;\middle|\; \mathcal{A} \text{ is a } \sigma\text{-algebra over } \Omega \text{ and } \mathcal{C} \subset \mathcal{A}\right\},$$

since the intersection of σ-algebras is again a σ-algebra. This is referred to as the *σ-algebra generated by $\mathcal{C}$*. It is clearly the smallest σ-algebra that contains $\mathcal{C}$.

To be able to define a sequence of random variables, we require a product σ-algebra on a space of sequences. We define the σ-algebra over both a finite and a countable product of spaces.

Definition 2.4 (Product σ-algebra)**. Finite case:** For $n \in \mathbb{N}$, let $(\Omega_1, \mathcal{A}_1), \ldots, (\Omega_n, \mathcal{A}_n)$ be measurable spaces and let

$$\mathcal{C} = \{\prod_{i=1}^{n} A_i \mid A_i \in \mathcal{A}_i \text{ for each } i \in \{1, 2, \ldots, n\}\}.$$

The σ-algebra over $\prod_{i=1}^{n} \Omega_i$, denoted by $\bigotimes_{i=1}^{n} \mathcal{A}_i$, is called the *product σ-algebra* and is defined by

$$\bigotimes_{i=1}^{n} \mathcal{A}_i = \sigma(\mathcal{C}).$$

This makes $(\prod_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} \mathcal{A}_i)$ into a measurable space.

Countable case: Let $((\Omega_i, \mathcal{A}_i))_{i=1}^{\infty}$ be a sequence of measurable spaces and define the collection of *cylinder sets*

$$\mathcal{D} = \{\prod_{i=1}^{n} A_i \;\times \prod_{i=n+1}^{\infty} \Omega_i \mid n \in \mathbb{N},\; A_i \in \mathcal{A}_i \text{ for each } i \in \{1, 2, \dots, n\}\}.$$

The σ-algebra over $\prod_{i=1}^{\infty} \Omega_i$, denoted by $\bigotimes_{i=1}^{\infty} \mathcal{A}_i$, is called the *infinite-product σ-algebra* and is defined by

$$\bigotimes_{i=1}^{\infty} \mathcal{A}_i = \sigma(\mathcal{D}).$$

This makes $(\prod_{i=1}^{\infty} \Omega_i, \bigotimes_{i=1}^{\infty} \mathcal{A}_i)$ into a measurable space.

Remark 2.5. There is an alternative definition of the product σ-algebra that is equivalent to the one given in definition 2.4 that is particularly useful when constructing independent random variables. For the countable case, let $\Omega = \prod_{i=1}^{\infty} \Omega_i$ and

$$\pi_i : \Omega \to \Omega_i$$

be defined by

$$\pi_i(\omega_1, \omega_2, \dots) = \omega_i$$

which is the *projection map* onto the i-th coordinate. Define also

$$\mathcal{G} = \{\pi_i^{-1}(A_i) \mid A_i \in \mathcal{A}_i, i \in \mathbb{N}\},$$

consisting of preimages of measurable sets under the projection maps . Then the product σ-algebra is defined by

$$\bigotimes_{i=1}^{\infty} \mathcal{A}_i = \sigma(\mathcal{G})$$

which is the smallest σ-algebra on the countable product space such that all coordinate projections are measurable.

Example 2.6. a) For $n \geq 1$, let $\mathcal{T}$ be the collection of open sets of $\mathbb{R}$. Then the smallest σ-algebra on $\mathbb{R}$ that is generated by $\mathcal{T}$, that is, $\sigma(\mathcal{T})$, is called the *Borel σ-algebra* on $\mathbb{R}$ and is usually denoted by $\mathcal{B}(\mathbb{R})$.

b) The Borel σ-algebra over $\prod_{i=1}^{n} \mathbb{R} = \mathbb{R}^n$ is $\bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{R})$ which is denoted by $\mathcal{B}(\mathbb{R}^n)$.

c) In a similar manner, the Borel σ-algebra over the space of real sequences $\prod_{i=1}^{\infty} \mathbb{R} = \mathbb{R}^{\mathbb{N}}$ is $\bigotimes_{i=1}^{\infty} \mathcal{B}(\mathbb{R})$ and denoted by $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$.

From now on, these spaces will always be equipped with the Borel σ-algebras.

Definition 2.7 (Sub σ-algebra)**.** Let $\mathcal{A}$ and $\mathcal{F}$ be σ-algebras on a set Ω. Then $\mathcal{F}$ is a *sub σ-algebra* of $\mathcal{A}$ if $\mathcal{F} \subset \mathcal{A}$.

Definition 2.8 (Measure)**.** Let Ω be a set and $\mathcal{A}$ a σ-algebra over Ω. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a *measure* on $(\Omega, \mathcal{A})$ if the following conditions hold:

(i) for all $A \in \mathcal{A}$, $\mu(A) \geq 0$,

(ii) $\mu(\emptyset) = 0$,

(iii) for all $(A_i)_{i=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{A}$,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If μ is a measure on a σ-algebra $\mathcal{A}$ over Ω, the triple $(\Omega, \mathcal{A}, \mu)$ is said to be a *measure space*. We will later be restricting ourselves to a special case of a measure space, the *probability space*, which is just a measure space with $\mu(\Omega) = 1$ (where we restrict the codomain to [0,1]).

Definition 2.9 (σ-finite measure)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. The measure μ is called *σ-finite* if there exists a sequence of measurable sets $(A_i)_{i=1}^{\infty}$ such that

$$\Omega = \bigcup_{i=1}^{\infty} A_i$$

and $\mu(A_i) < \infty$ for each i.

Example 2.10. If $(\Omega, \mathcal{A}, \mu)$ is a probability space, then it is σ-finite. Indeed, let $A_i = \Omega$ for all $i \in \mathbb{N}$. Clearly the union is equal to Ω and $\mu(A_i) = \mu(\Omega) = 1 < \infty$.

The concept of measure spaces can be generalized to *product measure spaces* as follows: Let $(\Omega_1, \mathcal{A}_1, \mu_1), \dots, (\Omega_n, \mathcal{A}_n, \mu_n)$ be measure spaces with σ-finite measures, and let the product space $\Omega = \prod_{i=1}^n \Omega_i$ be endowed with $\mathcal{A} = \bigotimes_{i=1}^n \mathcal{A}_i$ as in definition 2.4. It can be shown that there exists a unique measure $\mu : \mathcal{A} \to [0, \infty]$, often denoted $\mu_1 \times \mu_2 \times \cdots \times \mu_n$, such that

$$\mu(A_1 \times A_2 \times \cdots \times A_n) = \prod_{i=1}^n \mu_i(A_i).$$

This construction defines a measure space $(\Omega, \mathcal{A}, \mu)$. We now define what it means for a (real-valued) function to be measurable.

Definition 2.11 (Measurable function)**.** Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is said to be *measurable* (or $\mathcal{A}_1$-measurable) if $f^{-1}(A)$ is a measurable set for all $A \in \mathcal{A}_2$.

Remark 2.12. Note that if f is $\mathcal{F}$-measurable and $\mathcal{F}$ is a sub σ-algebra of $\mathcal{A}$, then f is also $\mathcal{A}$-measurable.

The following function bear som important properties in Ergodic Theory. It will be used to deduce the strong law of large numbers. The first thing we need to do is to show that it is measurable.

Example 2.13. The *shift map* $\varphi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ defined by

$$\varphi(x_1, x_2, \dots) = (x_2, x_3, \dots).$$

is measurable. To see this, recall that the Borel σ-algebra $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$ is generated by sets of the form $C = \{(x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} | \ (x_{i_1}, x_{i_2}, \dots, x_{i_n}) \in B\}$ where $B \in \mathcal{B}(\mathbb{R}^n)$, so we only need to show measurability for any such cylinder set. Note that

$$\varphi^{-1}(C) = \{(x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} \mid (x_{i_1+1}, x_{i_2+1}, \dots, x_{i_n+1}) \in B\}.$$

which is also a cylinder set and thus belongs to $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$.

In measure theory, we are often concerned only with properties that hold on sets of positive measure. This perspective allows us to simplify many arguments by ignoring sets of measure zero. For example, some important convergence theorems do not hold unless the convergence fails only on a set of measure zero. The following definition makes this idea more precise.

Definition 2.14. Let $\mathcal{P}$ be a property concerning the points ω of a measure space $(\Omega, \mathcal{A}, \mu)$. Then $\mathcal{P}$ is said to be *true almost everywhere* (a.e.) if the set of points for which $\mathcal{P}$ is not true has measure zero.

Remark 2.15. If the measure space is a probability space, the term "almost everywhere" (a.e.) is often referred to as "almost surely" (a.s.).

Given a measurable function function from an arbitrary measure space, the measure can be pushed forward along the function to obtain a new measure on the codomain. This is commonly used in probability spaces to obtain the so called probability distribution of a random variable, which we will introduce later.

Proposition 2.16. *Let $(\Omega_1, \mathcal{A}_1, \mu)$ be a measure space and $(\Omega_2, \mathcal{A}_2)$ a measurable space. Let $f : \Omega_1 \to \Omega_2$ be a measurable function. Then the mapping*

$$\mu_f : \mathcal{A}_2 \to [0, \infty]$$

given by

$$\mu_f(A) = \mu(f^{-1}(A))$$

is a measure on $\mathcal{A}_2$. The measure μ_f is usually referred to as the measure induced by f.

Proof. Clearly $\mu_f \geq 0$ since $\mu \geq 0$, and $\mu_f(\emptyset) = \mu(f^{-1}(\emptyset)) = \mu(\emptyset) = 0$. Given a sequence $(A_i)_{i=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{A}_2$, we have

$$\begin{aligned}
\mu_f\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(f^{-1}\left(\bigcup_{i=1}^{\infty} A_i\right)\right) \\
&= \mu\left(\bigcup_{i=1}^{\infty} f^{-1}(A_i)\right) \\
&= \sum_{i=1}^{\infty} \mu\left(f^{-1}(A_i)\right) \\
&= \sum_{i=1}^{\infty} \mu_f(A_i),
\end{aligned}$$

so it is indeed a measure on $\mathcal{A}_2$. $\square$

2.2 Lebesgue Integration

We briefly go through the construction of the Lebesgue integral. A main advantage of the Lebesgue integral over the Riemann integral, especially in probability theory, is that it allows, for example, expectations of both discrete and continuous random variables to be treated simultaneously. Other important advantages include its convergence properties that follows from the Bounded Convergence Theorem, which will be used later in the text. The statements are given without further discussion.

Definition 2.17 (Indicator Function)**.** Let Ω be a set and $A \subset \Omega$. The indicator function of A, denoted $\mathbf{1}_A : \Omega \to \mathbb{R}$, is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

Proposition 2.18. *The indicator function $\mathbf{1}_A : \Omega \to \mathbb{R}$ on a measurable space $(\Omega, \mathcal{A})$ is measurable if and only if A is measurable.*

Definition 2.19. (Simple Function) Let $(\Omega, \mathcal{A})$ be a measurable space. A function $f : \Omega \to \mathbb{R}$ is called a simple function if it can be written as

$$f(\omega) = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}(\omega),$$

where $c_i \in \mathbb{R}$ and A_i are measurable sets.

Remark 2.20. Note that since the sum of measurable functions are measurable, it follows that every simple function is measurable.

Definition 2.21. A simple function f on a measure space $(\Omega, \mathcal{A}, \mu)$, given by

$$f(\omega) = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}(\omega)$$

is said to be integrable if $\mu(A_i) < \infty$ for all i for which $c_i \neq 0$. We define the integral of f with respect to μ to be

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

We adopt the convention that $c_i \mu(A_i) = 0$ whenever $c_i = 0$ and $\mu(A_i) = \infty$.

We now extend the definition of the integral to more general functions. The same notational convention will be used.

Definition 2.22. If $f : \Omega \to [0, \infty)$ is a measurable function on a measure space $(\Omega, \mathcal{A}, \mu)$, its integral is defined as

$$\int f \, d\mu = \sup \left\{ \int g \, d\mu : 0 \leq g \leq f, \; g \text{ simple} \right\}.$$

Finally, A measurable function $f : \Omega \to \mathbb{R}$ is said to be *integrable* (w.r.t. the measure μ) if

$$\int |f| \, d\mu < \infty.$$

If f is integrable, its *integral* is defined by

$$\int f \, d\mu = \int \max(f, 0) \, d\mu - \int \max(-f, 0) \, d\mu.$$

If A is a measurable set and f is integrable, then the *integral over A* is defined by

$$\int_A f \, d\mu = \int f \cdot \mathbf{1}_A \; d\mu.$$

Remark 2.23. If the integral is taken over the entire set Ω, we will sometimes omit writing it out.

Bear in mind the following proposition, which will be the key to deducing the probabilistic version of the Ergodic Theorem from the general one.

Proposition 2.24 (Change of variable formula)**.** *Let $(\Omega_1, \mathcal{A}_1, \mu)$ be a measure space and $(\Omega_2, \mathcal{A}_2)$ a measurable space. Let $g : \Omega_1 \to \Omega_2$ be a measurable function and $f : \Omega_2 \to \mathbb{R}$ integrable w.r.t the measure induced by g. Then for any $A \in \mathcal{A}$ we have*

$$\int_A f \, d\mu_g = \int_{g^{-1}(A)} f \circ g \, d\mu.$$

3 Random Variables

We start by reviewing the definition of a probability space.

Definition 3.1 (Probability space)**.** A probability space is a measure space $(\Omega, \mathcal{A}, P)$, where Ω is a set called the sample space, $\mathcal{A}$ is a σ-algebra called the event space, and P is a measure, called the probability measure, that satisfies $P(\Omega) = 1$.

From now on, we will denote a general probability space by $(\Omega, \mathcal{A}, P)$.

3.1 Definition and Properties

Loosely speaking, a random variable is a function that assigns each element of the sample space to a real number. Because the probability measure is defined only on measurable sets, the function itself must be measurable

Definition 3.2 (Random variable)**.** A *random variable* is a measurable function $X : \Omega \to \mathbb{R}$.

More generally, if $X_1, ..., X_n : \Omega \to \mathbb{R}$ are random variables, a *random vector* is a measurable function $(X_1, ..., X_n) : \Omega \to \mathbb{R}^n$.

We have previously defined the measure induced by a measurable function between arbitrary measurable spaces (see proposition 2.16). In the probabilistic setting, the measure *induced by a random vector* $Y = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$, consistent with the previous notation, is given by

$$P_Y : \mathcal{B}(\mathbb{R}^n) \to [0, 1], \quad P_Y(A) = P(Y^{-1}(A))$$

and is called the probability distribution of Y.

Two events $A_1, A_2 \in \mathcal{A}$ are said to be independent if $P(A_1 \cap A_2) = P(A_1)P(A_2)$. This notion can be extended to more than two events by induction. Since for any random variable X, the preimage of any borel set belongs to $\mathcal{A}$, independence of n random variables is naturally defined as follows:

Definition 3.3. The random variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are said to be *independent* if

$$P\left(\bigcap_{i=1}^{n} X_i^{-1}(B_i)\right) = \prod_{i=1}^{n} P_{X_i}(B_i)$$

for all $B_i \in \mathcal{B}(\mathbb{R})$.

Proposition 3.4. *Suppose $X_1, X_2, \dots, X_n$ are independent random variables with distributions $P_{X_1}, \dots, P_{X_n}$ respectively. Then the random vector $(X_1, \dots, X_n)$ has distribution $P_{X_1} \times \cdots \times P_{X_n}$.*

Proof. Using the product measure constructed earlier, for $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ we have that

$$\begin{aligned} P\big((X_1, \dots, X_n)^{-1}(B_1 \times \cdots \times B_n)\big) &= P\big(X_1^{-1}(B_1) \cap \cdots \cap X_n^{-1}(B_n)\big) \\ &\overset{\text{indep.}}{=} \prod_{i=1}^{n} P_{X_i}(B_i) \\ &= P_{X_1} \times \cdots \times P_{X_n}(B_1 \times \cdots \times B_n). \end{aligned}$$

□

3.2 Expectations

As claimed earlier, one important advantage of measure-theoretic probability theory is that it allows us to define objects such as the expected value for both discrete and continuous random variables simultaneously, since, we can use the counting measure for discrete random variables and the Lebesgue measure for the continuous case.

Definition 3.5 (Expected Value)**.** Let X be a random variable defined on $(\Omega, \mathcal{A}, P)$. The *expected value of X*, denoted by $E[X]$, is the defined as

$$E[X] = \int_{\Omega} X \, dP.$$

We will sometimes write $E_P[X]$ to emphasize the underlying probability measure.

The probabilistic Ergodic Theorem is formulated in terms of conditional expectation, and therefore we will introduce it here.

Remark 3.6. Since a random variable X is, in particular, a measurable function, it follows by the definition given in section 2.2 that it is integrable if $E[|X|] < \infty$.

Definition 3.7 (Conditional Expectation)**.** Let $\mathcal{F} \subset \mathcal{A}$ be a sub σ-algebra, and $X : \Omega \to \mathbb{R}$ a ($\mathcal{A}$-measurable) random variable with $E[|X|] < \infty$.

The *conditional expectation* of X given $\mathcal{F}$, written as $E[X \mid \mathcal{F}]$, is an $\mathcal{F}$-measurable random variable that satisfies

$$\int_F E[X \mid \mathcal{F}]\, dP = \int_F X\, dP$$

for all $F \in \mathcal{F}$.

Lemma 3.8. *The conditional expectation $E[X \mid \mathcal{F}]$ given in definition 3.7 is integrable, i.e., $E[E[X \mid \mathcal{F}]] < \infty$.*

Proof. Let $A = \{\omega \in \Omega \mid E[X \mid \mathcal{F}](\omega) > 0\}$. It is easy to verify that A is $\mathcal{F}$-measurable. By the monotonicity of the integral, We have

$$\int_A E[X \mid \mathcal{F}]\, dP = \int_A X\, dP \leq \int_A |X|\, dP.$$

We also know that $\Omega \setminus A$ is $\mathcal{F}$-measurable and again by monotonicity we get

$$\int_{\Omega \setminus A} -E[X \mid \mathcal{F}]\, dP = \int_{\Omega \setminus A} -X\, dP \leq \int_{\Omega \setminus A} |X|\, dP.$$

Using the fact that $|E[X \mid \mathcal{F}]| = E[X \mid \mathcal{F}]$ on A and $|E[X \mid \mathcal{F}]| = -E[X \mid \mathcal{F}]$ on $\Omega \setminus A$, we conclude that $E|E[X \mid \mathcal{F}]| \leq E|X|$. □

Theorem 3.9 (Kolmogorov's Existence and Uniqueness Theorem for Conditional Expectation)**.** *The conditional expectation $E[X \mid \mathcal{F}]$ exists and is unique almost surely.*

Proof. See for example [1, pp. 206–207]. □

Example 3.10. If $\mathcal{F}$ is the trivial σ-algebra (see example 2.2) , Then $E[X \mid \mathcal{F}] = E[X]$. Indeed, we have that

$$\int_\emptyset E[X \mid \mathcal{F}]\, dP = \int_\emptyset X\, dP = 0$$

and

$$\int_\Omega E[X \mid \mathcal{F}]\, dP = \int_\Omega X\, dP = E[X].$$

4 Stochastic Processes and Extension of measures

We begin by defining what a (general) stochastic process is.

Definition 4.1. Let $\mathbb{T}$ be an index set (often interpreted as time). A *stochastic process* is an ordered collection of random variables $(X_t : \Omega \to \mathbb{R} \mid t \in \mathbb{T})$.

4.1 Sequences of Random Variables

If $\mathbb{T} = \mathbb{N}$, we obtain a sequence of random variables $X_1, X_2, \ldots$ The process is then called a discrete-time stochastic process. We will only consider sequences of random variables. It will be essential for stating and proving the Strong Law of Large Numbers.

A stochastic process can also be viewed as a measurable function $Y = (X_1, X_2, \ldots) : \Omega \to \mathbb{R}^{\mathbb{N}}$. This is helpful because it allows us to study the process as one mathematical object.

Proposition 4.2 (Measurability of The Canonical Stochastic Process)**.** *The mapping*

$$Y = (X_1, X_2, \ldots) : \Omega \to \mathbb{R}^{\mathbb{N}}$$

defined by

$$Y(\omega) = (X_1(\omega), X_2(\omega), \ldots),$$

is a measurable function. Hence it defines a stochastic process.

We now present the three most important properties that a sequence of random variables can satisfy.

Definition 4.3. The random variables $X_1, X_2, ...$ are said to be independent if every finite subcollection $(X_{i_1}, \ldots, X_{i_n})$ consists of independent random variables.

Definition 4.4. A sequence of random variables $X_1, X_2, \ldots$ are said to be *identically distributed* if $P_{X_1} = P_{X_n}$ for all $n \geq 1$.

Definition 4.5. A sequene of random variables $X_1, X_2, \ldots$ is said to be *stationary* if for any $k \geq 0$ and any $n \geq 0$, the random vectors $(X_1, X_2, ..., X_n)$ and $(X_{k+1}, X_{k+2}, ..., X_{k+n})$ has the same distribution.

Example 4.6. Let $X_1, X_2, \dots$ be i.i.d. random variables. Then they form a stationary sequence. Given $k, n \geq 0$, define the random vector

$$Y_k = (X_{k+1}, X_{k+2}, \dots, X_{k+n}).$$

Using Proposition 3.4 and the fact that the variables have the same distribution, we obtain for $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ that

$$P_{Y_0}(B_1 \times \cdots \times B_n) \overset{\text{indep.}}{=} P_{X_1}(B_1) \times \cdots \times P_{X_n}(B_n) \overset{\text{i.d.}}{=} P_{X_1}^{\otimes n}(B_1 \times \cdots \times B_n).$$

Similarly, for $B_{k+1}, \dots, B_{k+n} \in \mathcal{B}(\mathbb{R})$,

$$P_{Y_k}(B_{k+1} \times \cdots \times B_{k+n}) \overset{\text{indep.}}{=} P_{X_{k+1}}(B_{k+1}) \times \cdots \times P_{X_{k+n}}(B_{k+n}) \overset{\text{i.d.}}{=} P_{X_{k+1}}^{\otimes n}(B_{k+1} \times \cdots \times B_{k+n}).$$

Since X_1 and X_{k+1} are identically distributed, it follows that the distributions of Y_0 and Y_k agree. Hence, the sequence is stationary.

4.2 Kolmogorov's 0-1 Law and Extension Theorem

Theorem 4.7 (Kolmogorov's Extension Theorem)**.** *Suppose we are given a probability measures $P_1, \dots, P_n$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R})), \dots, (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ respectively, which are consistent, that is,*

$$P_{n+1}(B_1 \times \cdots \times B_n \times \mathbb{R}) = P_n(B_1 \times \cdots \times B_n)$$

for all $B_1, ..., B_n \in \mathcal{B}(\mathbb{R})$. Then there exists a unique probability measure $P^{\mathbb{N}} : \mathcal{B}(\mathbb{R}^{\mathbb{N}}) \to [0, 1]$ such that

$$P^{\mathbb{N}}(B_1 \times \cdots \times B_n \times \mathbb{R}^{\mathbb{N}}) = P_n(B_1 \times \cdots \times B_n).$$

Proof. A detailed proof can be found in [1, Appendix A.3, pp. 464–466] □

There is an important special case of the theorem. If P is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $P_n = P^{\otimes n}$, which is the product measure with respect to itself, then by the definition of the product measure (see the discussion on page 8) and the fact that the measure on the whole space is 1, we have

$$\begin{aligned} P^{\otimes(n+1)}(B_1 \times \cdots \times B_n \times \mathbb{R}) &= P^{\otimes n}(B_1 \times \cdots \times B_n) \cdot P(\mathbb{R}) \\ &= P^{\otimes n}(B_1 \times \cdots \times B_n). \end{aligned}$$

This shows consistency. Hence by the extension theorem, the measure can be uniquely extended to $P^{\mathbb{N}}$ such that

$$\begin{aligned} P^{\mathbb{N}}\left(B_1 \times \cdots \times B_n \times \mathbb{R}^{\mathbb{N}}\right) &= P^{\otimes n}\left(B_1 \times \cdots \times B_n\right) \\ &= \prod_{i=1}^{n} P(B_i). \end{aligned}$$

Note how, given that a sequence of random variables $X_1, X_2, \dots$ has the same distribution P_{X_1}, we can let $P = P_{X_1}$ to imply independence.
Before we state Kolmogorov's $0-1$ law, we simplify notations and state an important definition. Let $X_1, X_2, \dots$ be a sequence of random variables on $(\Omega, \mathcal{A}, P)$ and write

$$\sigma(X_k, X_{k+1}, \dots) = \sigma(X_k^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})).$$

This is the smallest σ-algebra that makes all the random variables in the sequence measurable. Define the *tail σ-algebra*

$$\mathcal{T} = \bigcap_{k=1}^{\infty} \sigma(X_k, X_{k+1}, \dots).$$

Note that $\mathcal{T}$ is a sub σ-algebra of $\mathcal{A}$.

Theorem 4.8 (Kolmogorov's 0-1 Law)**.** *Suppose that $X_1, X_2, \dots$ are independent random variables. If $A \in \mathcal{T}$, then $P(A) \in \{0, 1\}$.*

Proof. The proof is short and straightforward and can be found in [1, pp. 81–82]. □

5 Birkhoff's Ergodic Theorem

5.1 Ergodicity and Measure-preserving

Throughout this section, let $(\Omega, \mathcal{A}, \mu)$ a be probability space.

Definition 5.1 (Measure-Preserving)**.** A measurable map $T : \Omega \to \Omega$ is said to be *measure-preserving* (or to preserve μ) if $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{A}$.

From now on, we let T be a measure-preserving transformation with respect to $(\Omega, \mathcal{A}, \mu)$. We call the quadruple $(\Omega, \mathcal{A}, \mu, T)$ a *measure-preserving system*. A set $A \in \mathcal{A}$ is said to be T-*invariant* if $T^{-1}(A) = A \quad a.s.$. We will adopt the convention that two sets are considered equal if they are equal up to a set of measure 0, and will sometimes omit writing "a.s." explicitly.

Definition 5.2 (Ergodicity)**.** The transformation T is *ergodic* if every T-invariant set has measure 0 or 1.

5.2 Statement of the Ergodic Theorem

Theorem 5.3 (Birkhoff)**.** *Let $(\Omega, \mathcal{A}, \mu, T)$ be a measure-preserving system and suppose that $f : \Omega \to \mathbb{R}$ is integrable, then the limit*

$$f^*(\omega) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(\omega))$$

exists a.e. Moreover, f^* is integrable, and

$$\int f^* \, d\mu = \int f \, d\mu.$$

If T is ergodic, then

$$f^*(\omega) = \int f \, d\mu \quad a.e.$$

5.3 A Probabilistic Formulation

Fix a measure-preserving system $(\Omega, \mathcal{A}, P, T)$. Define $\mathcal{I}$ to be the collection of T-invariant events, i.e.,

$$\mathcal{I} = \{A \in \mathcal{A} : T^{-1}(A) = A\}.$$

The following lemma shows that it is a σ-algebra.

Lemma 5.4. *The collection $\mathcal{I}$ is a sub σ-algebra of $\mathcal{A}$.*

Proof. First, note that $\Omega \in \mathcal{I}$ Since T is a map from Ω to Ω:

$$\begin{aligned} T^{-1}(\Omega) &= \{\omega \in \Omega \mid T(\omega) \in \Omega\} \\ &= \Omega. \end{aligned}$$

Now let $A \in \mathcal{I}$, then since $T^{-1}(A) = A$ we have

$$T^{-1}(A^c) = T^{-1}(A)^c = A^c$$

hence $A^c \in \mathcal{I}$. Finally, we show that $\mathcal{I}$ is closed under countable unions. If $A_i \in \mathcal{I}$ for $i \in \mathbb{N}$, then $T^{-1}(A_i) = A_i$ for each $i \in \mathbb{N}$ so

$$T^{-1}(\bigcup_{i=1}^{\infty} A_i) = \bigcup_{i=1}^{\infty} T^{-1}(A_i) = \bigcup_{i=1}^{\infty} A_i.$$

□

Recall that T is ergodic if every T-invariant subset has measure 0 or 1. In other words, T is ergodic if for every $A \in \mathcal{I}$, $P(A) \in \{0, 1\}$, which is equivalent to saying that $\mathcal{I}$ is trivial.

Lemma 5.5. *A random variable $X : \Omega \to \mathbb{R}$ is $\mathcal{I}$-measurable if and only if it is T-invariant, i.e., $X \circ T = X$ a.s.*

Proof. Suppose X is $\mathcal{I}$-measurable. Then for any $B \in \mathcal{B}(\mathbb{R})$ we have $X^{-1}(B) \in \mathcal{I}$. By the definition of $\mathcal{I}$,

$$T^{-1}(X^{-1}(B)) = X^{-1}(B) \quad a.s..$$

But

$$T^{-1}(X^{-1}(B)) = \{\omega : T(\omega) \in X^{-1}(B)\} = \{\omega : X(T(\omega)) \in B\} = (X \circ T)^{-1}(B).$$

Thus

$$(X \circ T)^{-1}(B) = X^{-1}(B) \quad a.s..$$

This shows that

$$X \circ T = X \quad a.s..$$

Conversely, suppose $X \circ T = X$ a.s. . Let $B \in \mathcal{B}(\mathbb{R})$. We need to show that $X^{-1}(B) \in \mathcal{I}$, i.e.

$$T^{-1}(X^{-1}(B)) = X^{-1}(B) \quad a.s..$$

But

$$T^{-1}(X^{-1}(B)) = (X \circ T)^{-1}(B).$$

Since $X \circ T = X$ a.s.,

$$(X \circ T)^{-1}(B) = X^{-1}(B).$$

Thus $T^{-1}(X^{-1}(B)) = X^{-1}(B)$ a.s., so $X^{-1}(B) \in \mathcal{I}$.
Hence X is $\mathcal{I}$-measurable.

$\square$

Theorem 5.6 (Birkhoff - A Probabilistic Formulation)**.** *Let $(\Omega, \mathcal{A}, P, T)$ be a measure-preserving system, and let $X : \Omega \to \mathbb{R}$ be a random variable with $E[|X|] < \infty$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X(T^j(\omega)) = E[X \mid \mathcal{I}](\omega) \quad a.s..$$

Furthermore, if T is ergodic, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X(T^j(\omega)) = E[X] \quad a.s..$$

We aim to show that this probabilistic formulation is equivalent to the one given in theorem 5.3. Suppose theorem 5.3 holds and let $f = X$. Since X is, in particular, a measurable function defined on a probability space, the assumptions are fulfilled. Now, let $\mathcal{I}$ be as above. To conclude that $f^* = E[X \mid \mathcal{I}]$ *a.s.*, we has to show that f^* is an $\mathcal{I}$-measurable random variable that satisfies

$$\int_I E[X \mid \mathcal{I}]\, dP = \int_I X\, dP$$

for all $I \in \mathcal{I}$. Define

$$f_n(w) = \frac{1}{n} \sum_{j=0}^{n-1} X(T^j(\omega)).$$

Both X and T are measurable by assumption so the composition is also measurable. Furthermore, finite sums of measurable functions are measurable, and the pointwise limits are then measurable, which shows that f^* is a random variable. Moreover, f^* is T invariant:

$$\begin{aligned}
f^*(T(\omega)) &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} X(T^{j+1}(\omega)) \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} X(T^{j}(\omega)) \\
&= \lim_{n\to\infty} \left(\frac{1}{n}\sum_{j=0}^{n-1} X(T^{j}(\omega)) - \frac{1}{n}X(\omega) + \frac{1}{n}X(T^n(\omega)) \right) = \\
&= f^*(\omega) \quad a.s.
\end{aligned}$$

since $X(\omega) \in \mathbb{R}$ and $X(T^n(\omega)) \in \mathbb{R}$. Hence it follows by lemma 5.5 that f^* is $\mathcal{I}$-measurable. Let $I \in \mathcal{I}$. Using the change of variable formula (proposition 2.24) and the fact that I is invariant and T is measure-preserving we have

$$\begin{aligned}
\int_I f_n \, dP_T &= \int_I \left(\frac{1}{n}\sum_{j=0}^{n-1} X \circ T^j \right) dP_T \\
&= \frac{1}{n}\sum_{j=0}^{n-1} \int_I X \circ T^j \, dP_T \\
&= \frac{1}{n}\sum_{j=0}^{n-1} \int_I X \, d(P \circ T^{-j}) \\
&= \frac{1}{n}\sum_{j=0}^{n-1} \int_I X \, dP \\
&= \int_I X \, dP
\end{aligned}$$

But note also that

$$\begin{aligned}
\int_I f_n \, dP_T &= \int_I f_n \, d(P \circ T^{-1}) \\
&= \int_I f_n \, dP
\end{aligned}$$

thus we have

$$\int_I f_n \, dP = \int_I X \, dP \quad \text{for all } n.$$

Finally, since f_n converges pointwise to f^*, we may, omitting the details of finding a dominating function, apply the Bounded Convergence Theorem (see [3, pp. 54–56] for a full proof and statement)

$$\begin{aligned}\int_I f^* \, dP &= \int_I \lim_{n\to\infty} f_n \, dP \\ &= \lim_{n\to\infty} \int_I f_n \, dP \\ &= \int_I X \, dP.\end{aligned}$$

The claim $f^* = E[X \mid \mathcal{I}]$ follows from the uniqueness of conditional expectation (see theorem 3.9). The last part of the theorem is clear, since if T is ergodic, then $\mathcal{I}$ is trivial and the claim then follows from example 3.10

6 The Strong Law of Large Numbers

Start by observing that a sequence of random variables can be viewed as an element of $\mathbb{R}^{\mathbb{N}}$, namely $(X_1(\omega), X_2(\omega), \dots) \in \mathbb{R}^{\mathbb{N}}$. To deduce the Strong Law of Large number from the Ergodic Theorem, we need construct a measure-preserving system on the space of real sequences $\mathbb{R}^{\mathbb{N}}$. We start by showing some crucial properties about the shift map defined in example 2.13, where its measurability was shown.

Lemma 6.1. *The shift map is ergodic.*

Proof. Recall the shift map

$$\varphi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}, \quad \varphi(x_1, x_2, \dots) = (x_2, x_3, \dots),$$

and let $\mathcal{I}_\varphi$ be the collection of φ-invariant events. Let $A \in \mathcal{I}_\varphi$, then

$$A = \varphi^{-1}(A) = \{(x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} \mid (x_2, x_3, \dots) \in A\} \in \sigma(X_2, X_3, ...)$$

since A only depends on $(x_2, x_3, \dots)$ and not on x_1. Similarly,

$$A = \varphi^{-2}(A) = \{(x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} \mid (x_3, x_4, \dots) \in A\} \in \sigma(X_3, X_4, ...).$$

By induction we get that

$$A \in \bigcap_{k=1}^{\infty} \sigma(X_k, X_{k+1}, \dots) = \mathcal{T},$$

where $\mathcal{T}$ is the tail σ-algebra. This implies that $\mathcal{I}_\varphi \subset \mathcal{T}$. When the probability distribution $P_{X_1}^{\mathbb{N}}$ is put on $\mathbb{R}^{\mathbb{N}}$, we get that the sequence $X_1, X_2, \dots$ are independent by the extension theorem, so $\mathcal{T}$ is trivial by the 0-1 law. Thus $\mathcal{I}_\varphi$ is trivial which shows that $P(A) \in \{0, 1\}$.

□

Now let $X_1, X_2, \dots$ be a sequence of random variables on $(\Omega, \mathcal{A}, P)$. By proposition 4.2, the random variables induces a stochastic process

$$Y = (X_1, X_2, \dots) : \Omega \to \mathbb{R}^{\mathbb{N}}$$

defined by

$$Y(\omega) = (X_1(\omega), X_2(\omega), \dots).$$

Since Y is a measurable function, we can endow $\mathbb{R}^{\mathbb{N}}$ with the induced measure which we refer to as the *distribution of the stochastic process* Y. We prove the following lemma:

Lemma 6.2. *If $X_1, X_2, \dots$ are i.i.d., then the shift map preserves the distribution P_Y.*

Proof. Note that the distribution of Y is given by

$$P_Y : \mathcal{B}(\mathbb{R}^{\mathbb{N}}) \to [0, 1], \quad P_Y(B) = P(Y^{-1}(B)).$$

We want to show that $P_Y(\varphi^{-1}(B)) = P_Y(B)$ for all $B \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$. Given $B \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$, we have

$$\begin{aligned} P_Y(\varphi^{-1}(B)) &= P(Y^{-1}(\varphi^{-1}(B))) \\ &= P((\varphi \circ Y)^{-1}(B)) \\ &= P_{\varphi \circ Y}(B). \end{aligned}$$

Now, since a sequence of *i.i.d.* random variables is stationary (see example 4.6), the distribution of the shifted sequence $\varphi \circ Y$ agrees with the distribution of Y, that is,

$$P_{\varphi \circ Y}(B) = P_Y(B),$$

which completes the proof of the lemma. □

Theorem 6.3 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables over $(\Omega, \mathcal{A}, P)$ with $\mathbb{E}[|X_1|] < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = E[X_1] \quad a.s..$$

Proof. Since $X_1, X_2, \dots$ are *i.i.d.*, lemma 6.2 says that the shift map φ preserves the distribution P_Y, the unique measure on $\mathbb{R}^{\mathbb{N}}$. Hence $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), P_Y, \varphi)$ is a measure-preserving system. Furthermore, φ is ergodic by lemma 6.1. Now let $X = \pi_1$, the projection map onto the first coordinate (which is measurable by construction), and $T = \varphi$ in Theorem 5.6. Thus, almost surely, we get that

$$\begin{aligned}
E_{P_Y}[X] = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X\big(T^j(w_1, w_2, \dots)\big) &= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X\big(T^{j-1}(w_1, w_2, \dots)\big) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \pi_1\big(\varphi^{j-1}(X_1(w), X_2(w), \dots)\big) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \pi_1\big(X_j(w), X_{j+1}(w), \dots\big) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(w).
\end{aligned}$$

But since $X_j = \pi_1 \circ T^j$, we conclude that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(w) = E_p(X_1) \quad a.s.$$

$\square$

Note how this can be solved with theorem 5.3 using the change of variable formula (proposition 2.24):

$$E_{P_Y}[\pi_1] = \int_{\mathbb{R}^{\mathbb{N}}} \pi_1 \, dP_Y \; = \; \int_{\Omega} \pi_1 \circ Y \, dP = E_P[X_1].$$

But this is essentially already done in the discussion about the probabilistic formulation of the Ergodic Theorem.

References

[1] Rick Durrett. *Probability: Theory and Examples*. 5th edition, Cambridge University Press, 2019.

[2] Wikipedia. *Measure theory*. `https://en.wikipedia.org/wiki/Measure_theory`.

[3] Avner Friedman. *Foundations of Modern Analysis*. Dover Publications, 1982.

[4] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. 2nd edition, Wiley, 1999.

[5] M. M. Rao. *Stochastic Processes: General Theory*. Mathematics and Its Applications, Vol. 157, Springer, 1992.

[6] Daniel Ahlberg. Personal communication. Universitetslektor, Stockholms universitet, 2025.