Probability Theory II

— Exam Preparation —

Heinrich-Gregor Zirnstein

December 3, 2015

Contents

1 Preliminaries

This section recalls some very useful notions and results from probability theory.

Definition. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The **expectation value** or **mean** $E[X]$ of the random variable is defined as the integral

$$E[X] := \int_\Omega X(\omega)\, dP(\omega),$$

provided that the random variable is integrable, $X \in L^1(\Omega, \mathcal{A}, P)$. Often, we denote the latter condition with "$E[|X|] < \infty$".

Definition. Let X be a random variable that is square integrable, $X \in L^2$. Then, its **variance** is defined as

$$\boxed{V[X] := E\big[(X - E[X])^2\big] = E\big[X^2\big] - E[X]^2\,.}$$

Remark. Often, we say that the variable has "finite variance", $V[|X|] < \infty$, and mean that it is square integrable. Note that this already implies that it has finite mean as well.

The mother of all probabilistic inequalities is probably the following:

Lemma 1.1 (**Markov's inequality**). *Let X be a random variable with finite mean, $E[|X|] < \infty$. Then, for $\delta > 0$, we have*

$$\boxed{P\left(|X| \geq \delta\right) \leq \frac{E[|X|]}{\delta}.}$$

Proof.

$$P\left(|X| \geq \delta\right) = \int_{\{|X|\geq\delta\}} \mathbb{1}\, dP \leq \frac{1}{\delta} \int_{\{|X|\geq\delta\}} \delta\, dP \leq \frac{1}{\delta} \int_{\{|X|\geq\delta\}} |X|\, dP = \frac{E[|X|]}{\delta}.$$

□

A slight generalization is given by

Lemma 1.2 (**Markov's inequality, general form**). *Let $X : \Omega \to \mathbb{R}$ be a random variable and $\varphi : (0, \infty) \to (0, \infty)$ be a* monotonically increasing *measurable function. If the random variable $\varphi \circ |X|$ is integrable, the we have*

$$\boxed{P(|X| \geq \delta) \leq \frac{E[\varphi \circ |X|]}{\varphi(\delta)}.}$$

Proof. By Markov's inequality, $P\left(|X| \geq \delta\right) = P\left(\varphi \circ |X| \geq \varphi(\delta)\right) \leq \frac{1}{\varphi(\delta)} E[\varphi \circ |X|]$. □

Corollary 1.3 (**Chebychev's inequality**). *For a random variable X with finite variance, $X \in L^2$, we have*

$$\boxed{P(|X - E[X]| \geq \delta) \leq \frac{V[X]}{\delta^2}.}$$

Proof. Apply Markov's inequality to $\varphi(x) = x^2$. □

Sometimes, we want to prove that a random variable has finite mean. The following criterion is very useful for that

Lemma 1.4 (**Integrability of a random variable**)**.**

$$\boxed{E[|X|] < \infty \iff \sum_{n=1}^{\infty} P(|X| > n) < \infty.}$$

Proof. "$\Longrightarrow$" Applying Markov's inequality directly is not enough to conclude the convergence of the series on the right. But we can consider the double sum

$$\sum_{n=1}^{N} P\left(|X| > n\right) = \sum_{n=1}^{N} \int_{\Omega} \mathbb{1}_{\{|X|>n\}}\, dP \leq \int_{\Omega} \sum_{n=1}^{[|X|]} \mathbb{1}\, dP \leq \int_{\Omega} |X|\, dP = E[|X|] < \infty$$

where $[x]$ denotes the largest integer smaller than the real number x. Taking the limit $N \to \infty$ gives the desired result.

"$\Longleftarrow$" This is essentially the same double sum, read from right to left. We have

$$\int_{|X|<C} |X|\, dP \leq \int_{|X|<C} \sum_{n=1}^{[|X|]+1} \mathbb{1}\, dP \leq 1 + \sum_{n=1}^{\infty} \int_{|X|<C} \mathbb{1}_{\{|X|>n\}}\, dP < 1 + \sum_{n=1}^{\infty} P\left(|X| > n\right).$$

Letting $C \to \infty$ gives the desired result.s □

2 Convergence I – The Law of Large Numbers

2.1 Introduction

The **law of large numbers** is based on the following observation: If we repeatedly throw a six-sided die, then the *average* of the face values over many throws will tend to 3.5. This is equal to the *expected value*, where we argue that each side has a probability of 1/6 to show up, so we expect an average of $(1+2+3+4+5+6)/6 = 3.5$.

More generally, we can consider any sequence $(X_n)_{n=1}^{\infty}$ of independent identically distributed random variables. This means that we are repeating a single experiment X_i independently many, many times. Each experiment has the (same) expected value $E[X_n] = \mu$, and we intuitively expect that the average of this sequence will tend towards the expected value

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu.$$

This is the law of large numbers.

The goal of the following sections is to prove this statement and to make it more precise: In which sense does this sequence convergence? What is the rate of convergence?

In this section, we will establish the weak and strong form of the law of large numbers, which say that under certain conditions, the averages converge in probability, resp. almost surely. In the next section, Section 3, we will show that the rate of convergence is, in some sense, universal. This is the central limit theorem.

2.2 Independence

Definition. Two events A, B are called **independent** if

$$P(A \cap B) = P(A)P(B).$$

Definition. A finite collection of events $(A_1, \ldots, A_n)$ is called **mutually independent** if, for any subset of indices $\{n_1, n_2, \ldots, n_k\} \subset \{1, \ldots, n\}$, we have

$$P\left(A_{n_1} \cap A_{n_2} \cap \ldots \cap A_{n_k}\right) = P(A_{n_1})P(A_{n_2}) \cdots P(A_{n_k}).$$

In other words, events are mutually independent if each event is independent from any *intersection* of the other events.

Remark. Note that in the definition of mutual independence, some of the events may be the same, e.g. $A_1 = A_2 = A$. It makes sense to say that an event is independent of itself, which simply means that $P(A \cap A) = P(A)P(A)$, or $P(A) = 0$ or 1.

So far, we have captured the notion of independence of individual events. For instance, given a red and a blue die, it is fair to say that the events "the red die rolls a 1" and "the blue die rolls a 4" are independent. However, we also want to capture the notation that *any* possible roll of the red die is independent of the rolls of the blue die.

Definition. Let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ be collections of sets. We say that these collections are **mutually independent** if, for any subset of indices $\{n_1, n_2, \ldots, n_k\} \subset \{1, \ldots, n\}$, we have

$$\boxed{P\left(A_{n_1} \cap A_{n_2} \cap \ldots \cap A_{n_k}\right) = P(A_{n_1})P(A_{n_2}) \cdots P(A_{n_k}) \quad \text{for all sets } A_{n_j} \in \mathcal{E}_{n_j}.}$$

For example, the can let $\mathcal{E}_1$ be the possible outcomes of the red die and $\mathcal{E}_2$ be the possible outcomes of the blue die. We will often apply this notion to the case where the collections $\mathcal{E}_k$ are σ-algebras.

Definition. An infinite family of collections of sets, $(\mathcal{E}_j)_{j\in J}$ is said to be **independent** if every finite subfamily is mutually independent.

Definition. A family of random variables $(X_j)_{j\in J}$ is called **independent** if the corresponding family of generated σ-algebras $(X_j^{-1}(\mathcal{B}_{\mathbb{R}}))_{j\in J}$ is independent.

Example. Let $((\Omega_k, \mathcal{A}_k, P_k))_{k=1}^n$ be a finite collection of measure spaces. Consider the product space $(\Omega, \mathcal{A}, P) = (\prod_{k=1}^n \Omega_k, \otimes_{k=1}^n \mathcal{A}_k, \otimes_{k=1}^n P_k)$ and let $\pi_k : \Omega \to \Omega_k$ denote the projection onto the k-th coordinate. Then, the family $(\pi_k^{-1}(\mathcal{A}_k))_{k=1}^n$ is independent.

Example. Two random variables $X, Y : \Omega \to \mathbb{R}$ induce a joint probability distribution on $\mathbb{R}^2$, namely the distribution of the pair (X, Y). The two variables are independent if and only if their joint distribution is a product measure on $\mathbb{R}^2$.

Proof. By the previous example, if the joint distribution is the product measure, then the random variables are independent. On the other hand, if they are independent, then let P denote the joint probability distribution and P_X and P_Y the marginal distributions on $\mathbb{R}^2$. By assumption, we have $P(A \times B) = P_X(A)P_Y(B)$ for any product of Borel sets $A \times B \subset \mathcal{B}^2$. The collection of finite disjoint unions of such product sets form an algebra that generates the product σ-algebra. By the Hahn-Kolmogorov extension theorem A.4, the measure P has a unique extension to the whole σ-algebra. But by the previous equation, it coincides with the product measure. □

Lemma 2.1. *If two random variables $X, Y \in L^1$ are* independent*, then their product has the expectation value*

$$E[XY] = E[X]\, E[Y]\,.$$

Proof. By the previous example, this is a direct consequence of Fubini's theorem. □

Lemma 2.2. *If two random variables $X, Y \in L^2$ are* independent*, then their variance is additive*

$$V[X + Y] = V[X] + V[Y].$$

Proof. Without loss of generality, we may assume that $E[X] = E[Y] = 0$. By the previous lemma, we have

$$\begin{aligned} V[X+Y] = E\big[(X+Y)^2\big] &= E\big[X^2\big] + 2E[XY] + E\big[Y^2\big] \\ &= E\big[X^2\big] + 2\underbrace{E[X]\,E[Y]}_{=0} + E\big[Y^2\big] = V[X] + V[Y]. \end{aligned}$$

□

Notation. Often, we will be concerned with sequences of random variables $(X_n)_{n=1}^{\infty}$ that are independent and identically distributed. We will call this an **i.i.d. sequence**.

Remark. It is not immediately obvious that infinite sequence of independent variables exist at all. It requires the construction of a product measure on the product σ-algebra. We will consider this question in more detail in Section 5, where we will prove the Kolmogorov consistency theorem 5.1 and the very general Ionescu-Tulcea theorem 5.2.

Example. An i.i.d. sequence of random variables $(X_n)_{n=1}^{\infty}$ whose values may only be 1 or 0 is also called a **Bernoulli process**. Each of the mutually independent variables is equal to 0 with probability p and equal to 1 with probability $(1-p)$. When $p = 1/2$, this corresponds to an infinite sequence of coin flips, where 1 corresponds to "head" and 0 corresponds to "tail".

Definition. Let $X_1, X_2, \dots$ be a sequence of *independent* random variables $X_n : \Omega \to \mathbb{R}$. Let $\mathcal{F}_n$ be the σ-algebra generated by the random variables starting from index n, that is the variables $X_n, X_{n+1}, \dots$. Then, the **tail σ-algebra** is the σ-algebra

$$\mathcal{F}_\infty = \bigcap_{n=1}^{\infty} \mathcal{F}_n.$$

Any event $A \in \mathcal{F}_\infty$ is called a **tail event**.

Example. The events

$$\{\omega \in \Omega : \text{ the sequence } X_n(\omega) \text{ is bounded}\} \text{ and } \left\{\omega \in \Omega : \limsup_{n\to\infty} X_n(\omega) = 1\right\}$$

are tail events, because they do not depend on the values of the first few variables. On the other hand, the event $\{\omega \in \Omega : \sup_{n\in\mathbb{N}} |X_n(\omega)| \leq 1\}$ is *not* a tail event.

It turns out that tail events are rather deterministic:

Proposition 2.3 (**Kolmogorov's zero-one law**). *Let $X_1, X_2, \dots$ be a sequence of independent variables. Then, any tail event $A \in \mathcal{F}_\infty$ occurs either almost surely or almost never, that is $P(A) = 1$ or $P(A) = 0$.*

The proof of this statement is not obvious, we will need the following lemma whose proof uses Dynkin's theorem A.1:

Lemma 2.4 (**Independence of σ-algebras**). *Let $(\mathcal{E}_j)_{j\in J}$ be an independent family, such that each collection $\mathcal{E}_j$ is closed under finite intersections. Then, the family of o-algebras $(\sigma(\mathcal{E}_j))_{j\in J}$ is independent as well.*

Proof. We can restrict our attention to finite families. For simplicity, we only consider the case of a family with two members $\mathcal{E}_1, \mathcal{E}_2$. The general case is similar and follows by induction.

Consider the collection

$$\mathcal{A} = \{A \subset \Omega : P(A \cap B) = P(A)P(B) \text{ for all } B \in \mathcal{E}_2\}$$

of events that are independent from the sets in the second collection $\mathcal{E}_2$. By definition, we have $\mathcal{E}_1 \subset \mathcal{A}$. We show that $\mathcal{A}$ is a Dynkin system:

1. Clearly, $\in \mathcal{A}$ and $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$, then we have $P(A^c \cap B) = 1 - P(A \cap B^c) = 1 - P((A \cap B) \uplus B^c) = P(B) - P(A \cap B) = P(A^c)P(B)$. Hence, the complement A^c is also in the collection $\mathcal{A}$.
3. If $A_1, A_2, \dots$ are a sequence of disjoint subsets in $\mathcal{A}$, then thanks to the σ-additivity of the probability measure, their union is also in $\mathcal{A}$.

Hence, by Dynkin's theorem, the set $\mathcal{A}$ also contains the σ-algebra $\sigma(\mathcal{E}_1)$.

Now, we consider the collection

$$\mathcal{B} = \{B \subset \Omega : P(A \cap B) = P(A)P(B) \text{ for all } A \in \sigma(\mathcal{E}_1)\}$$

of events that are independent from the σ-algebra $\sigma(\mathcal{E}_1)$. By construction, of the collection $\mathcal{A}$, we have $\mathcal{E}_2 \subset \mathcal{B}$. By essentially the same calculations as above, we conclude that this is a Dynkin system, and we obtain that $\sigma(\mathcal{E}_1)$ is independent of $\sigma(\mathcal{E}_2)$ as desired. □

Proof of Kolmogorov's zero-one law. We show that tail events are independent of themselves.

Let $\mathcal{A}_n$ denote the σ-algebra generated by the variable X_n. Moreover, let

$$\mathcal{B}_n = \{A \subset \Omega : \ A = A_1 \cap A_2 \cap \ldots \cap A_n \text{ with } A_k \in \mathcal{A}_k\}$$

denote the collection of sets that can be written as finite intersections of events from the first n σ-algebras. Complementarily, let

$$\mathcal{C}_n = \{A \subset \Omega : \ \exists m \in \mathbb{N}, A = A_{n+1} \cap A_{n+2} \cap \ldots \cap A_m \text{ with } A_k \in \mathcal{A}_k\}$$

be the collection of finite intersections of events from the σ-algebras starting at the index $n+1$.

By assumption, the collections $\mathcal{B}_n$ and $\mathcal{C}_n$ are independent for all $n \in \mathbb{N}$. By the previous lemma, the collections $\mathcal{B}_n$ and $\sigma(\mathcal{C}_n) = \mathcal{F}_n$ are independent as well. In particular, the collections $\mathcal{B}_n$ and $\mathcal{F}_\infty$ are independent. This implies the independence of the collections $\bigcup_{n=1}^\infty \mathcal{B}_n$ and $\mathcal{F}_\infty$. Applying the lemma again, we see that the σ-algebras $\sigma(\bigcup_{n=1}^\infty \mathcal{B}_n) = \mathcal{F}_1$ and $\mathcal{F}_\infty$ are independent. The desired result follows. □

2.3 Notions of convergence of random variables

Definition. Let $(X_n)_{n=1}^\infty$ be a sequence of random variables. We say that the sequence converges to a random variable X ...

1. ... **in probability** if for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

2. ... **in the p-th mean** ($p \geq 1$) if all random variables have p-th moments, $X_n \in L^p(\Omega, \mathcal{A}, P)$, and the sequence converges in L^p,

$$\lim_{n\to\infty} \int_\Omega |X_n - X|^p \, dP = 0.$$

From the theory of L^p spaces, we know that $X \in L^p$ in this case as well.

3. ... **almost surely** if

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 0.$$

In other words, the pointwise limit of the sequence is almost surely equal to X.

Reminder. For any probability space, we have a continuous embedding $L^p(\Omega, \mathcal{A}, P) \subset L^{p'}(\Omega, \mathcal{A}, P)$ as long as $p > p'$. In other words, if a random variable has p-th moments, then it also has p'-th moments for any $p' < p$.

Lemma 2.5 (**Notions of convergence**). *We have the following implications between these notions of convergence:*

(a) Convergence almost surely $\implies$ convergence in probability.

(b) Convergence in the p-th mean $\implies$ convergence in probability.

(c) If there exists a random variable $Y \in \mathcal{L}^p$ with $|X_n| \leq Y$, then convergence almost surely implies convergence in the p-th mean.

Proof. On (a). For $\varepsilon > 0$, consider the indicator function $Y_n = \mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}$. This function is integrable with $|Y_n| \leq 1$. For any point ω, the convergence $X_n(\omega) \to X(\omega)$ implies that $Y_n(\omega) \to 0$. Hence, we have $Y_n \to 0$ almost surely. By Lebesgue's dominated convergence theorem, it follows that

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = \lim_{n\to\infty} E[Y_n] = E[0] = 0.$$

On (b). By Markov's inequality,

$$P\left(|X_n - X| \geq \varepsilon\right) \leq \frac{E[|X_n - X|^p]}{\varepsilon^p}.$$

By assumption, this tends to zero as $n \to \infty$.

On (c). From $|X_n|^p \leq Y^p$, we conclude that $|X|^p \leq Y^p$ as well. Hence, the p-th power of the difference is dominated by an integrable function, $|X_n - X|^p \leq (|X_n| + |X|)^p \leq (2Y)^p$. Moreover, we have $|X_n - X|^p \to 0$ almost everywhere and Lebesgue's dominated convergence theorem implies that $X_n \to X$ in the p-the mean. □

Example. "Escape to vertical infinity." Consider the unit interval with the Lebesgue measure as a probability space $\Omega = [0, 1]$. The sequence of random variables $X_n = n\mathbb{1}_{[0,1/n]}$ converges almost surely and in probability to the zero function, but not in the p-th mean, because $E[|X_n|] = 1$.

Example. "Typewriter sequence". Consider again the unit interval. The sequence of random variables

$$X_{2^k+n} = \mathbb{1}_{[n2^{-k},(n+1)2^{-k}]} \quad \text{for } k, n \in \mathbb{N},\ 0 \leq n < 2^k$$

converges in the p-th mean and hence in probability, because their supports become thinner and thinner. However, it does not converge anywhere pointwise.

We will show more implications between these differing notions of convergence in Section 2.5 where we introduce the simple but useful Borel-Cantelli lemma.

2.4 Weak Law of Large Numbers

The **weak law of large numbers** states that under very mild assumptions, the average converges to the expected value *in probability*.

The statement has an easy proof if the random variables have a *finite* second moment.

Proposition 2.6 (**Weak law of large numbers**, L^2)**.** *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables in L^2 with $E[X_n] = \mu$ and $V[X_n] = \sigma^2$. Then, we have*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \text{ in probability.}$$

Proof. Let $Y_n = (X_1 + \cdots + X_n)/n$ denote the average. Then, we can use Chebychev's inequality and independence to estimate

$$P\left(|Y_n - \mu| > \delta\right) \leq \frac{V[Y_n]}{\delta^2} = \frac{1}{\delta^2}\left(\sum_{k=1}^{n} V\left[\frac{X_k}{n}\right]\right) = \frac{1}{\delta^2} n \frac{\sigma^2}{n^2} = \frac{1}{n}\frac{\sigma^2}{\delta^2} \to 0$$

for $n \to \infty$. □

Remark. Actually, the assumption that the random variables be *identically* distributed is not needed in this case.

But the statement can be strengthened: It also holds if the random variables have a finite expectation value $E[|X_k|] < \infty$, the existence of the second moment is not needed.

Proposition 2.7 (**Weak law of large numbers**, L^1)**.** *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables in L^1 with $E[X_n] = \mu$. Then, we have*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \text{ in probability.}$$

Proof. For a random variable X and a constant $C > 0$, let

$$X^C(\omega) := \begin{cases} X(\omega), & \text{if } |X(\omega)| \leq C \\ 0, & \text{otherwise.} \end{cases}$$

denote the truncated random variable. It has finite moments of any order.

To show convergence in probability, fix $\delta > 0$ and consider an arbitrary $\varepsilon > 0$ with $\delta > \varepsilon$. By assumption, the random variables are in L^1 and identically distributed, so we can find a constant $C > 0$ such that

$$E[|X_k^C - X_k|] = \int_{\{|X_k| \geq C\}} |X_k| dP \leq \varepsilon$$

for all indices $k \in \mathbb{N}$. If we denote the expectation values with $\mu = E[X_k]$ and $\mu^C = E[X_k^C]$, then this implies

$$|\mu - \mu^C| \leq \varepsilon < \delta.$$

But it also implies

$$E\left[\left|\frac{X_1^C + X_2^C + \cdots + X_n^C}{n} - \frac{X_1 + X_2 + \cdots + X_n}{n}\right|\right] \leq n\frac{1}{n}\varepsilon = \varepsilon,$$

independent of the number n. Using Markov's inequality, we can estimate

$$\begin{aligned} P\left(\left|\frac{X_1 \cdots + X_n}{n} - \mu\right| > 3\delta\right) &\leq P\left(\left|\frac{X_1^C + \cdots + X_n^C}{n} - \frac{X_1 + \cdots + X_n}{n}\right| > \delta\right) \\ &\quad + P\left(\left|\frac{X_1^C + \cdots + X_n^C}{n} - \mu^C\right| > \delta\right) + P\left(\left|\mu - \mu^C\right| > \delta\right) \\ &\leq \frac{\varepsilon}{\delta} + P\left(\left|\frac{X_1^C + \cdots + X_n^C}{n} - \mu^C\right| > \delta\right) + 0. \end{aligned}$$

Now, the random variables X_k^C have finite second moment, hence we can apply the previous version of the weak law of large numbers, Proposition 2.6, to see that the second summand on the right-hand side tends to zero as $n \to \infty$. Since $\varepsilon > 0$ was arbitrary, we see that the whole right-hand side also tends to zero in the limit $n \to \infty$, as desired. □

Example. If the probability distribution of the random variables does not have a finite mean, then the weak law may or may not hold.

For instance, the Cauchy distribution $p(x)dx = \frac{1}{\pi}\frac{1}{1+x^2}dx$ does not have a finite mean. It can be shown that if $(X_n)_{n=1}^{\infty}$ is an i.i.d. sequence of random variables whose distribution is the Cauchy distribution, then the average $(X_1 + \cdots + X_n)/n$ also has the Cauchy distribution. Hence, the sequence of averages does not converge to a constant value in probability. (Using the method of characteristic functions, Section 3.2, we can calculate the characteristic function of the variable X_n as $\phi(t) = e^{-|t|}$ and obtain the characteristic function of the average as $\phi(t/n)^n = \phi(t)$, as desired.)

2.5 The Borel-Cantelli Lemma

A very simple but useful tool for establishing convergence almost surely is the **Borel-Cantelli lemma**.

Reminder. Let $(A_n)_{n=1}^{\infty}$ be a sequence of sets. Then, the set

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k$$

("approximation from above") denotes the set of points that are contained in infinitely many sets A_n. Furthermore, the set

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k$$

("approximation from below") denotes the set of points that are contained in all but finitely many sets A_n. These two notions correspond to the limes superior and limes inferior of the corresponding characteristic functions.

Remark. The complements satisfy

$$\left(\limsup_{n\to\infty} A_n\right)^c = \liminf_{n\to\infty} A_n^c.$$

Lemma 2.8 (**1st Borel-Cantelli lemma**). *Let $(A_n)_{n=1}^{\infty}$ be a sequence of events. Then, the following implication holds*

$$\boxed{\sum_{n=1}^{\infty} P(A_n) < \infty \implies P\left(\limsup_{n\to\infty} A_n\right) = 0}$$

A partial converse is given by the

Lemma 2.9 (**2nd Borel-Cantelli lemma**). *Let $(A_n)_{n=1}^{\infty}$ be a sequence of events. Then, the following implication holds*

$$\boxed{\text{all } A_n \text{ independent } \wedge \sum_{n=1}^{\infty} P(A_n) = \infty \implies P\left(\limsup_{n\to\infty} A_n\right) = 1}$$

Proof. On 1. We have

$$P\left(\limsup_{n\to\infty} A_n\right) = P\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \leq P\left(\bigcup_{k=N}^{\infty} A_k\right) \leq \sum_{k=N}^{\infty} P\left(A_k\right)$$

for any index $N > 0$. In the limit $N \to \infty$, the right-hand side tends to zero.

On 2. Consider the probability of the complementary event,

$$1 = P\left(\left(\limsup_{n\to\infty} A_n\right)^c\right) = P\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c\right) \leq \sum_{n=1}^{\infty} P\left(\bigcap_{k=n}^{\infty} A_k^c\right).$$

We can use independence and the divergence of the series the to show that each summand vanishes by estimating

$$\begin{aligned} P\left(\bigcap_{k=n}^{\infty} A_k^c\right) &= \lim_{N\to\infty}\prod_{k=n}^{N} P(A_k^c) = \lim_{N\to\infty}\prod_{k=n}^{N}(1 - P(A_k)) \\ &\leq \lim_{N\to\infty}\prod_{k=n}^{N} e^{-P(A_k)} = \lim_{N\to\infty} e^{-\sum_{k=n}^{N} P(A_k)} = 0. \end{aligned}$$

Here, we have used the estimate $1 - x \leq e^{-x}$ for $x \geq 0$. □

Remark. One corollary of the second Borel-Cantelli lemma is the following: Imagine a random experiment with two outcomes, "success" and "failure". If we repeat this experiment independently infinitely often, then no matter how small the probability of success, as long as it is non-zero, we will eventually have success. For instance, if a monkey presses the keys of a typewriter randomly and independently infinitely often, then it will, at some point, have written Shakespeare's "Hamlet" at least once. (In fact, infinitely often).

Proof. One particular edition of the play "Hamlet" consists of $l = 173305$ letters. For each number $k \in \mathbb{N}$, let A_k be the event that the $(lk+1)$-th letter to the $l(k+1)$-th letter of the monkey's writing match the play exactly. These events are independent, and their probability is very small but non-zero, $P(A_k) = 26^{-l}$. Hence, the series $\sum_{k=1}^{\infty} P(A_k)$ diverges and by the second Borel-Cantelli lemma, infinitely many of them will occur with probability 1. □

We know that L^p spaces are complete. Something similar holds for convergence in probability:

Lemma 2.10 (**Completeness, in probability**). *If a sequence of random variables $(X_n)_{n=1}^{\infty}$ is a **Cauchy sequence in probability**,*

$$\forall \varepsilon, \delta > 0, \exists N, \forall n, m \geq N, \quad P(|X_n - X_m| > \varepsilon) < \delta,$$

then there exists a limit such that $X_n \to X$ in probability.

Proof. We use the trick of making sets exponentially small. By assumption, for each $k \in \mathbb{N}$, we can find a subsequence n_k such that

$$P\left(|X_{n_{k+1}} - X_{n_k}| > 2^{-k}\right) < 2^{-k}.$$

By the Borel-Cantelli lemma 2.8, we conclude that

$$P\left(\limsup_{k\to\infty} |X_{n_{k+1}} - X_{n_k}| > 2^{-k}\right) = 0.$$

But this means that the series

$$X_{n_k} = X_{n_0} + \sum_{k=1}^{\infty} (X_{n_{k+1}} - X_{n_k})$$

converges absolutely almost everywhere. In other words, the subsequence converges pointwise almost surely, $X_{n_k} \to X$.

Moreover, we have the triangle inequality

$$P(|X_m - X| > \varepsilon) \leq P\left(|X_m - X_{n_k}| > \frac{\varepsilon}{2}\right) + P\left(|X_{n_k} - X| > \frac{\varepsilon}{2}\right).$$

Since the convergence $X_{n_k} \to X$ is also in probability, we can make second term small enough by choosing k large. The first term can be made small enough by using the Cauchy condition and making k even larger and m larger than n_k. Hence, the original sequence converges in probability. □

Examining the proof again, we see that we have also established the following:

Corollary 2.11. *If a sequence $X_n \to X$ converges in probability, then there exists a* subsequence *such that $X_{n_j} \to X$ converges almost surely.*

If we strengthen the Cauchy condition to include a supremum, then the sequence of random variables will converge almost surely:

Lemma 2.12 (**Convergence almost surely from maximal inequality**). *Let $(Y_n)_{n=1}^\infty$ be a sequence of random variables such that*

$$\forall \varepsilon, \delta > 0, \exists N, \forall n, m \geq N,\ P\left(\sup_{m \leq k \leq n} |Y_k - Y_m| > \delta\right) \leq \varepsilon.$$

Then, the sequence converges to a limit $Y_n \to Y$ almost surely.

Proof. By assumption, we can find an increasing sequence of indices $(m_j)_{j=1}^\infty$ such that

$$P\left(\sup_{m_j \leq k} |Y_k - Y_{m_j}| > \delta\right) \leq 2^{-j}.$$

By the Borel-Cantelli lemma 2.8, this means that

$$P\left(\limsup_{j \to \infty} \left\{\sup_{m_j \leq k} |Y_k - Y_{m_j}| > \delta\right\}\right) = 0.$$

Taking the union over $\delta = 1/l$ with natural numbers $l \in \mathbb{N}$, we see that the event

$$\bigcup_{l=1}^\infty \left\{\omega \in \Omega : \sup_{m_j \leq k} |Y_k(\omega) - Y_{m_j}(\omega)| > \frac{1}{l} \text{ for infinitely many } j \in \mathbb{N}\right\}$$

has probability zero. But since the sequence m_j is increasing, this is precisely the set of points ω where the sequence $(Y_k(\omega))_{k=1}^\infty$ is *not* a Cauchy sequence. □

2.6 Strong Law of Large Numbers I

We now want to strengthen the weak law of large numbers and investigate whether the average over independent experiments converges almost surely, which is stronger than convergence in probability. This will be the **strong law of large numbers**, though in this section, we will only achieve a partial result.

Proposition 2.13 (**Strong Law of Large Numbers**, L^4). *Let $(X_n)_{n=1}^\infty$ be an i.i.d. sequence of random variables in L^4 with $E[X_n] = \mu$. Then, we have*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \text{ almost surely.}$$

Proof. Let $Y_n = (X_1 + \cdots + X_n)/n$ denote the average. Without loss of generality, we may assume that $E[X_k] = 0$. Then, we can use Markov's inequality with the fourth moment and obtain

$$P\left(|Y_n| > \delta\right) \leq \frac{1}{\delta^4} E\left[\left(\frac{X_1 + \cdots + X_n}{n}\right)^4\right] = \frac{1}{n^4 \delta^4} \sum_{i,j,k,l=1}^n E[X_i X_j X_k X_l]\,.$$

Since the variables are independent and we have assumed $E[X_k] = 0$, all terms in the sum that have one index distinct from the others must vanish. Hence, the right-hand side becomes equal to

$$P\left(|Y_n| > \delta\right) \leq \frac{1}{n^4 \delta^4} \left(n E\left[X_k^4\right] + n(n-1) E\left[X_k^2\right]^2\right) = \frac{1}{n^2 \delta^4} \left(E\left[X_k^2\right]^2 + O(1/n)\right).$$

Since the sum over inverse squares, $1/n^2$, converges, the sum over these probabilities converges as well. Applying the Borel-Cantelli Lemma 2.8, we see that

$$0 = P\left(\limsup_{n\to\infty} \{|Y_n| > \delta\}\right) = 1 - P\left(\left\{\limsup_{n\to\infty} |Y_n(\omega)| < \delta\right\}\right).$$

Since this holds for all $\delta > 0$, we conclude that $Y_n \to 0$ almost surely. □

2.7 Series of Independent Random Variables

In this section, we want to investigate the convergence of a series $\sum_{n=1}^{\infty} X_n$ of independent random variables. Ultimately, we want to use this to prove the strong law of large numbers in Section 2.8.

Let

$$S_n(\omega) := \sum_{k=1}^{n} X_n(\omega)$$

denote the n-th partial sum. The key lemma for estimating these sums is the following lemma named after Kolmogorov:

Lemma 2.14 (Kolmogorov's Maximal Inequality). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables with expectation $E[X_n] = 0$ and finite variance, $V[X_n] < \infty$. Then, we have*

$$\boxed{P\left(\sup_{1\le k\le n} |S_k(\omega)| > \delta\right) \le \frac{1}{\delta^2} V[S_n].} \tag{1}$$

Remark. Since the random variables are independent, the variance on the right-hand side is equal to $V[S_n] = V[X_1] + \ldots + V[X_n]$.

Remark. The bound $P(|S_n| > \delta) \le V[S_n]/\delta^2$ is a direct consequence of Markov's inequality, but Kolmogorov's inequality is stronger, because it bounds the *supremum* of the partial sums. Naively adding the weaker bounds would incur an additional factor of n.

Proof of Kolmogorov's maximal inequality. Consider the events

$$E_k := \{|S_1| \le \delta, |S_2| \le \delta, \ldots, |S_k| > \delta\}, \quad k = 1 \ldots n$$

where the first sums are smaller than the bound δ, but the k-th sum exceeds it. These events are mutually disjoint, and we have

$$\left\{\sup_{1\le k\le n} |S_k(\omega)| > \delta\right\} = \biguplus_{k=1}^{n} E_k.$$

For each of these events, we can use the Markov bound

$$P(E_k) = \int_{E_k} 1\, dP \le \frac{1}{\delta^2} \int_{E_k} S_k^2\, dP.$$

Now, the random variable $S_n - S_k$ is independent of the random variable $\chi_{E_k} S_k$, because the former only depends on the values of the variables $X_{k+1}, \ldots, X_n$, while the latter depends on the values of the variables $X_1, \ldots, X_k$. Hence, we can estimate the variance of the partial sum S_k on this set by

$$\begin{aligned}\int_{E_k} S_n^2 \, dP &= \int_{E_k} [(S_n - S_k)^2 + 2(S_n - S_k)S_k + S_k^2] \, dP \\ &= \int_{E_k} (S_n - S_k)^2 \, dP + 2 \underbrace{E[S_n - S_k]}_{=0} E[\chi_{E_k} S_k] + \int_{E_k} S_k^2 \, dP \\ &\geq \int_{E_k} S_k^2 \, dP.\end{aligned}$$

Summing over the disjoint events gives the desired inequality. □

The maximal inequality can be used to establish convergence almost surely by using Lemma 2.12 in the following way:

Proposition 2.15 (**Kolmogorov's One-Series Theorem**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of* independent *random variables in L^2 with mean $E[X_n] = 0$. Then, the following implication holds*

$$\sum_{n=1}^{\infty} V[X_n] < \infty \implies \sum_{n=1}^{\infty} X_n(\omega) \text{ converges almost surely.}$$

Proof. This is a direct application of Kolomogorov's maximal inequality 2.14. Namely, we have

$$P\left(\sup_{m \leq k \leq n} |S_k - S_m| > \delta \right) \leq \frac{1}{\delta^2} \sum_{k=m}^{n} V[X_k].$$

By assumption, there exists an index N such that the right-hand side is very small as soon as the indices $m, n \geq N$. □

Proposition 2.16 (**Kolmogorov's Two-Series Theorem**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of* independent *random variables in L^2. Then, the following implication holds*

$$\sum_{n=1}^{\infty} E[X_n] \text{ converges and } \sum_{n=1}^{\infty} V[X_n] < \infty \implies \sum_{n=1}^{\infty} X_n(\omega) \text{ converges almost surely.}$$

Proof. By Kolmogorov's One-Series Theorem, the sum of random variables $Y_n = X_n - E[X_n]$ converges. □

We now want to get rid of the assumption that the random variables are in L^2. As in the proof of the weak law of large numbers, Prop. 2.7, we consider the truncation of a random variable X. For any constant $C > 0$, let

$$X^C(\omega) := \begin{cases} X(\omega), & \text{if } |X(\omega)| \leq C \\ 0, & \text{otherwise.} \end{cases}$$

This variables has finite moments of any order.

Proposition 2.17 (**Kolmogorov's Three-Series Theorem**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of* independent *random variables without any condition on integrability. Then, the statement*

$$\sum_{n=1}^{\infty} X_n(\omega) \text{ converges almost surely}$$

holds if and only if there exists a constant $C > 0$ such that all of the following conditions are met

1. *The series $\sum_{n=1}^{\infty} P(|X_n| > C)$ converges.*
2. *The series $\sum_{n=1}^{\infty} E[X_n^C]$ converges.*
3. *The series $\sum_{n=1}^{\infty} V[X_n^C]$ converges.*

Proof. "$\Longleftarrow$" Applying the Borel-Cantelli lemma to the first condition, we obtain $P(\liminf_{n\to\infty}\{X_n \leq C\}) = 1$, which implies $P(\liminf_{n\to\infty}\{X_n = X_n^C\}) = 1$. In other words, the convergence of the series $\sum_{n=1}^{\infty} X_n(\omega)$ is equivalent to the convergence of the series $\sum_{n=1}^{\infty} X_n^C(\omega)$ with probability 1. But the convergence of the latter follows from Kolmogorov's Two-Series Theorem. □

The proof of the converse direction is more difficult. The essential difficulty is contained in the following lemma:

Lemma 2.18. *Let $(Y_n)_{n=1}^{\infty}$ be a sequence of independent random variables with mean $E[Y_n] = 0$ and bounded by $|Y_n| \leq C$. If the series $\sum_{n=1}^{\infty} Y_n$ converges almost surely, then the sum of variances $\sum_{n=1}^{\infty} V[Y_n]$ converges.*

Proof. Let $S_n = Y_1 + \ldots + Y_n$ denote the n-th partial sum. For a bound $l \geq 0$ and an index $n \in \mathbb{N}$, consider the event that the first n partial sums are smaller than the bound,

$$F_n := \{|S_1| \leq l, |S_2| \leq l, \ldots, |S_n| \leq l\}.$$

The event $F := \bigcup_{l=1}^{\infty}(\bigcap_{n=1}^{\infty} F_n)$ consists of all samples ω where the sequence $S_n(\omega)$ is bounded. Since this sequence converges almost surely by assumption, and every convergent sequence is bounded, this event has probability one. In particular, there must exist a bound $l > 0$ and a positive number $\delta > 0$ such $P(F_n) > \delta$ for all indices n.

Let $\sigma_n^2 = V[Y_n]$ denote the variance of the variable Y_n. Since the variable Y_n is independent of the variable $\mathbb{1}_{F_{n-1}} S_{n-1}$, we can write

$$\begin{aligned}
\int_{F_{n-1}} S_n^2 \, dP &= \int_{F_{n-1}} [S_{n-1}^2 + 2S_{n-1}Y_n + Y_n^2] \, dP \\
&= \int_{F_{n-1}} S_{n-1}^2 \, dP + 2E\big[\mathbb{1}_{F_{n-1}} S_{n-1}\big] \underbrace{E[Y_n]}_{=0} + P(F_{n-1})\sigma_n^2 \\
&\geq \int_{F_{n-1}} S_{n-1}^2 \, dP + \delta\sigma_n^2.
\end{aligned}$$

In other words, on the set F_{n-1}, the variance σ_n^2 contributes significantly to the square S_n^2.

On the other hand, we have

$$\begin{aligned}\int_{F_{n-1}} S_n^2\,dP &= \int_{F_n} S_n^2\,dP + \int_{F_n^c\cap F_{n-1}} S_n^2\,dP\\ &\le \int_{F_n} S_n^2\,dP + \int_{F_n^c\cap F_{n-1}} (l+C)^2\,dP\\ &= \int_{F_n} S_n^2\,dP + P(F_n^c\cap F_{n-1})(l+C)^2,\end{aligned}$$

that is the expected square S_n^2 on the set F_{n-1} is not that far away from its expectation on the set F_n.

Taking the sum over variances, we obtain

$$\delta\sum_{n=1}^N \sigma_n^2 \le \sum_{k=1}^N \left(\int_{F_k} S_k^2\,dP - \int_{F_{k-1}} S_{k-1}^2\,dP\right) + (l+C)^2\sum_{k=1}^N P(F_k^c\cap F_{k-1})$$

Since the sets $F_n^c\cap F_{n-1}$ are disjoint, this is a telescope sum, and we have

$$\delta\sum_{n=1}^N \sigma_n^2 \le \int_{F_N} S_N^2\,dP + (l+C)^2 \le l^2 + (l+C)^2.$$

Since $\delta > 0$, this gives the desired bound on the variances. □

Proof of Kolmogorov's Three-Series Theorem. "$\Longrightarrow$"

On 1. Since the series converges almost surely, for any constant $C > 0$, we have $|X_n(\omega)| \le C$ eventually with probability 1. Since the events are independent, we can apply the Borel-Cantelli lemma in reverse and obtain that the series of probabilities converges.

To prove the other two statements, we note that the convergence of the series implies that the series $\sum_{n=1}^\infty X_n^C$ converges as well. In other words, without loss of generality, we can assume that the variables are bounded, $|X_n| \le C$.

On 3. Draw an independent sequence $(X_n')_{n=1}^\infty$ from the same distribution. Then, the variables $Y_n = X_n - X_n'$ are bounded by $2C$, have mean 0, and the series $\sum_{n=1}^\infty Y_n$ converges. Moreover, due to independence, the variances satisfy $V[Y_n] = 2V[X_n]$. Applying the previous lemma gives the desired convergence.

On 2. Having established the convergence of the sum of variances, $\sum_{n=1}^\infty V[X_n] < \infty$, we can apply Kolomogorov's One-Series Theorem to conclude that the series $\sum_{n=1}^\infty (X_n - E[X_n])$ converges as well. Since the series $\sum_{n=1}^\infty X_n$ was assumed convergent, we conclude that the series of expectation values must converge as well. □

Example. The *harmonic series with random signs*,

$$\sum_{n=1}^\infty \pm\frac{1}{n},$$

where each sign is drawn independently and uniformly, converges almost surely.

2.8 Strong Law of Large Numbers II

In the previous section, we have considered infinite series of independent random variables. Now, we want to go back to the convergence of averages. The main technical tool for connecting series and averages is the following lemma:

Proposition 2.19 (**Kronecker's lemma**). *Let $0 < w_1 \leq w_2 \leq \dots$ be a sequence of positive weights that grow without bound, $\lim_{n\to\infty} w_n = \infty$. Then, for any sequence of real numbers, $(x_n)_{n=1}^{\infty}$, the following implication holds:*

$$\boxed{\sum_{n=1}^{\infty} x_n \text{ converges} \implies \lim_{n\to\infty} \frac{1}{w_n} \sum_{k=1}^{n} w_k x_k = 0}$$

Proof. Let $\varepsilon > 0$. By the Cauchy criterion, we can find an index $N > 0$ such that the partial sums satisfy $|\sum_{k=m}^{n} x_k| < \varepsilon$ whenever $m, n > N$. We split the sum on the right-hand side according to this index

$$y_n := \frac{1}{w_n} \sum_{k=1}^{n} w_k x_k = \frac{1}{w_n} \sum_{k=1}^{N} w_k x_k + \frac{1}{w_n} \sum_{k=N+1}^{n} w_k x_k.$$

For the second part, we use a telescope sum to estimate

$$\begin{aligned} \sum_{k=N+1}^{n} w_k x_k &= w_{N+1} \sum_{k=N+1}^{n} x_k + (w_{N+2} - w_{N+1}) \sum_{k=N+2}^{n} x_k + \dots + (w_n - w_{n-1}) x_N \\ \implies \left| \sum_{k=N+1}^{n} w_k x_k \right| &\leq w_{N+1}\varepsilon + (w_{N+2} - w_{N+1})\varepsilon + \dots + (w_n - w_{n-1})\varepsilon = w_n \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, we see that $y_n \to 0$ in the limit $n \to \infty$ as desired. □

Corollary 2.20. *If the series $X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3 + \dots$ converges almost surely, then the averages $(X_1 + \dots + X_n)/n$ tends to zero almost surely.*

Theorem 2.21 (**Strong Law of Large Numbers, L^1**). *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables in L^1 with $E[X_n] = 0$. Then, we have*

$$\frac{X_1 + X_2 + \dots + X_n}{n} \to 0 \text{ almost surely.}$$

Proof. We want to apply Kolmogorov's One-Series Theorem 2.15.

For each index n, consider the truncated random variable

$$Y_n(\omega) := \begin{cases} X_n(\omega), & \text{if } |X_n(\omega)| < n \\ 0, & \text{otherwise} \end{cases}$$

and let $\mu_n = E[Y_n]$ denote its expectation value. Since the original variables have the same distribution and are integrable, these expectation values converge to zero, $\lim_{n\to\infty} \mu_n = E[X_1] = 0$.

Consider the probability that the truncated and the original variables differ, $P(X_n \neq Y_n)$. Since the original variables are integrable, $E[|X_1|] < \infty$, we can apply Lemma 1.4 about integrability and obtain

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

Hence, by the Borel-Cantelli lemma 2.8, we can conclude that the sequences $X_n(\omega)$ and $Y_n(\omega)$ are eventually equal almost surely,

$$P(\{\omega \in \Omega : \exists N \in \mathbb{N},\ X_n(\omega) = Y_n(\omega) \text{ for all } n \geq N\}) = 1.$$

In other words, the series $\sum_{n=1}^{\infty} \frac{1}{n}(Y_n - \mu_n)$ converges almost everywhere if and only if the series $\sum_{n=1}^{\infty} \frac{1}{n}(X_n - \mu_n)$ does.

By construction, the random variables $\frac{1}{n}(Y_n - \mu_n)$ are uniformly bounded, independent and have vanishing expectation. To show that their infinite sum converges, we want to apply Kolmogorov's One-Series Theorem 2.15, which requires an estimate of their variances. If we let $d\alpha$ denote the probability distribution of the variable X_1, we have

$$\sum_{n=1}^{\infty} V\left[\frac{Y_n - \mu_n}{n}\right] = \sum_{n=1}^{\infty} \frac{1}{n^2} V\left[Y_n\right] = \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{|x|<n} |x|^2 \, d\alpha(x) = \int |x|^2 \sum_{n=1}^{\infty} \frac{1}{n^2} \mathbb{1}_{\{|x|<n|\}}(x) \, d\alpha(x).$$

The sum over inverse squares can be estimated by an integral

$$\sum_{n>|x|}^{\infty} \frac{1}{n^2} \leq \frac{1}{\lceil |x| \rceil^2} + \int_{y=\lceil |x| \rceil}^{\infty} \frac{1}{y^2} \, dy \leq \frac{2}{|x|},$$

and we obtain

$$\sum_{n=1}^{\infty} V\left[\frac{Y_n - \mu_n}{n}\right] \leq \int |x|^2 \frac{2}{|x|} \, d\alpha(x) = 2 \int |x| \, d\alpha(x) = 2E[|X_1|] < \infty.$$

Hence, the series $\sum_{n=1}^{\infty} \frac{1}{n}(Y_n - \mu_n)$ converges.

Now, we can apply Kronecker's lemma 2.19 to conclude that

$$\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{\mu_1 + \mu_2 + \cdots + \mu_n}{n} \to 0$$

almost surely. But since $\mu_n \to 0$, the second term tends to zero, and the desired result follows. □

Remark. If $X_1, X_2, \ldots$ is any sequence of random variables with the property that the averages converge almost surely, $\frac{1}{n}(X_1 + \ldots + X_n) \to Y$, then their limit is actually *constant* almost everywhere, $P(Y = c) = 1$ for some $c \in \mathbb{R}$. This follows from Kolmogorov's zero-one law 2.3: The limit variable Y is measurable with respect to the tail σ-algebra $\mathcal{F}_\infty$ and hence constant almost everywhere.

Remark. The tail σ-algebra $\mathcal{F}_\infty$ is defined without any reference to the probability distribution of the variables X_n, the same goes for the limit variable Y. However, the value c that the

limit is almost surely equal to may very well depend on the probability measures associated to the variables X_n.

Remark. It is also possible to show that the assumption of integrability, $X_n \in L^1$, cannot be dispensed with, as it follows the almost sure convergence of the averages:

Lemma 2.22 (**Strong law of large numbers requires L^1**). *Let $(X_k)_{k=1}^{\infty}$ be an i.i.d. sequence of random variables such that the averages $\frac{1}{n}(X_1 + \cdots + X_n)$ converge almost surely to a random variable Y. Then, the variables X_k are actually integrable, $X_k \in L^1$, and consequently $Y = E[X_k]$.*

Proof. To show that the random variables are integrable, we can use Lemma 1.4. For this, we have to show that the series $\sum_{n=1}^{\infty} P(|X_k| > n)$ converges. Since the variables are identically distributed, this is the same as showing $\sum_{n=1}^{\infty} P(|X_n| > n) < \infty$.

Let us assume the contrary, i.e. that the series diverges. Since the variables are independent, the second Borel-Cantelli Lemma 2.9 implies that

$$P\left(\limsup_{n\to\infty} \{|X_n| > n\}\right) = 1.$$

In other words, for almost all points $\omega \in \Omega$, infinitely many values satisfy $|X_n(\omega)| > n$. However, this is a contradiction to the convergence of the averages $Y_n = \frac{1}{n}(X_1 + \cdots + X_n)$, because

$$\frac{X_n}{n} = Y_n - \frac{n-1}{n} Y_{n-1} \to Y - Y = 0$$

almost surely. The desired conclusion follows. □

3 Convergence II – The Central Limit Theorem

3.1 Weak convergence of probability distributions

Previously, we had considered random variables $X_n : \Omega \to \mathbb{R}$ defined on an arbitrary probability space $(\Omega, \mathcal{A}, P)$, and discussed various notions of convergence of these functions. In the following, we will concentrate not so much on the domain Ω, but more on the target space $\mathbb{R}$. On this space, every random variable X_n induces a probability distribution $P_n = P \circ X_n^{-1}$ and we will consider convergence of these distributions.

Example. Let $(X_n)_{n=1}^\infty$ be an i.i.d. sequence of random variables and $(X'_n)_{n=1}^\infty$ be a sequence of identical random variables, $X'_n = X_1$. The corresponding sequences of probability distributions are constant and equal, but the former sequence of random variables does not converge in probability at all, whereas the latter one does so trivially.

Unlike a general probability space, the target space $\mathbb{R}$ has additional nice properties. For instance, it is a complete metric space.

Definition. A **polish metric space** is a complete metric space that is also separable.

Remark. A **polish space** is a topological space that can be given a complete, separable metric, but where this information is not part of the structure. In other words, there is no canonical choice of complete, separable metric.

Example. $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{C}^n$ are polish spaces. The $\ell_p(\mathbb{N})$ spaces are also polish spaces for $1 \leq p < \infty$, but $\ell_\infty(\mathbb{N})$ is not a polish metric space.

Every metric space is also a measurable space: the σ-algebra is chosen to be the Borel-σ-algebra.

Definition. Let $(P_n)_{n=1}^\infty$ be a sequence of probability measures on a *metric* space E. Then, we say that the sequence **converges weakly** to a probability distribution P if the integrals over all *bounded* continuous functions converge,

$$\boxed{P_n \Rightarrow P \;:\Longleftrightarrow\; \lim_{n\to\infty} \int f\, dP_n = \int f\, dP \quad \text{for all bounded continuous } f \in C_b(E, \mathbb{R}).}$$

Definition. A sequence $(X_n)_{n=1}^\infty$ of random variables **converges in distributions** to a random variable X if the corresponding probability distributions converge weakly, that is $P \circ X_n^{-1} \Rightarrow P \circ X^{-1}$.

Examples. 1. A sequence $(x_n)_{n=1}^\infty$ of points in E converges to a point $x \in E$ if and only if the corresponding sequence of Dirac measures converges weakly, $\delta_{x_n} \Rightarrow \delta_x$.

2. On the unit interval $E = [0, 1]$, the average of Dirac measures $\frac{1}{N}\sum_{n=1}^N \delta_{n/K}$ converges weakly to the uniform (Lebesgue) measure.

3. On the space $E = \mathbb{R}$, the sequence of Dirac measures $(\delta_n)_{n=1}^\infty$ does *not* converge weakly.

4. On the space $E = \mathbb{R}$, the sequence of uniform measures on the interval $[-n, n]$ does *not* converge weakly for $n \to \infty$.

Like any useful notion of convergence, the limit of a weakly converging sequence is unique, but we have to show this:

Lemma 3.1 (**Uniqueness** of the **weak limit**). *Let E be a metric space and let $(P_n)_{n=1}^{\infty}$ be a sequence of probability distributions on E. If P and $\tilde{P}$ are two probability measures such that $P_n \Rightarrow P$ and $P_n \Rightarrow \tilde{P}$, then the limits are equal, $P = \tilde{P}$.*

Proof. Consider the collection of sets A such that $P(A) = \tilde{P}(A)$. Apparently, this collection is a Dynkin system. If we can prove that it contains all closed sets, then, since the closed sets are stable under intersection, we can apply Dynkin's theorem and obtain that this collection contains all Borel sets.

Let A be a closed set. Let $f_\varepsilon(x)$ be a continuous function that approximates the indicator function $\mathbb{1}_A$ from above, for example $f_\varepsilon(x) = 1 - \min\{d(x, A)/\varepsilon, 1\}$. Weak convergence of the probability measures implies

$$\int f_\varepsilon \, dP = \lim_{n\to\infty} \int f_\varepsilon \, dP_n = \int f_\varepsilon \, d\tilde{P}.$$

In the limit $\varepsilon \to 0$, the monotone convergence theorem applies, and we see that the left- and right-hand sides converge to $P(A)$ and $\tilde{P}(A)$ respectively. □

We have defined weak convergence in terms of continuous functions, but it is also useful to understand weak convergence in terms of sets.

The main difficulty is the following example: Consider the real line. The sequence of Dirac measures $(\delta_{1/n})_{n=1}^{\infty}$ converges weakly to the Dirac measure δ_0. However, on the closed interval $A = [-1, 0]$, the measures satisfy $\delta_{1/n}(A) = 0$, whereas the limit satisfies $\delta_0(A) = 1$. In other words, mass can "move into" a closed set. Fortunately, that is the only thing that can happen: mass can never move out of a closed set, as the following lemma shows:

Proposition 3.2 (**Portemanteau-Theorem**). *Let $(P_n)_{n=1}^{\infty}$ be a sequence of probability measures and P be another probability measure on a metric space E. Then, the following are equivalent:*

1. *The sequence converges weakly, that is* $\lim_{n\to\infty} \int f \, dP_n = \int f \, dP$ *for each* bounded continuous *function $f \in C_b(E)$.*
2. *We have* $\lim_{n\to\infty} \int f \, dP_n = \int f \, dP$ *for every* bounded Lipschitz-continuous *function f.*
3. *We have* $\lim_{n\to\infty} \int f \, dP_n = \int f \, dP$ *for every* bounded measurable *function f where the set of discontinuities $\mathcal{U}_f = \{x \in E : f$ not continuous at $x\}$ has measure zero, $P(\mathcal{U}_f) = 0$.*
4. *For every open set U, we have* $\liminf_{n\to\infty} P_n(U) \geq P(U)$.
5. *For every closed set A, we have* $\limsup_{n\to\infty} P_n(A) \leq P(A)$.
6. *For every Borel set A with $P(\partial A) = 0$, we have* $\lim_{n\to\infty} P_n(A) = P(A)$.

Proof. The implications $3 \implies 1 \implies 2$ and $4 \iff 5$ are trivial. We only have to show the following implications:

"$4 \wedge 5 \implies 6$". Let A be a Borel set with $P(\partial A) = 0$. Then, the set $U = A \setminus \partial A$ is open and $\liminf_{n\to\infty} P_n(A) \geq \liminf_{n\to\infty} P_n(U) \geq P(U) = P(A)$. Likewise, the set $B = A \cup \partial A$ is closed, and we can conclude that $\limsup_{n\to\infty} P_n(A) \leq \limsup_{n\to\infty} P_n(B) \leq P(B) = P(A)$. Hence, the sequence $P_n(A)$ converges.

"$2 \implies 5$" Let A be a closed set. For $\varepsilon > 0$, consider the continuous functions $f_\varepsilon(x) = 1 - \min\{d(x, A)/\varepsilon, 1\}$. These function are Lipschitz-continuous and approximates the indicator function of the set A from above, which means that the functions are monotically decreasing in ε at every point x and that $\lim_{\varepsilon\to 0} f_\varepsilon(x) = \mathbb{1}_A(x)$. By assumption, for each $\varepsilon > 0$, we have

$$\limsup_{n\to\infty} P_n(A) \leq \limsup_{n\to\infty} \int f_\varepsilon \, dP_n = \int f_\varepsilon \, dP.$$

In the limit $\varepsilon \to 0$, the right-hand side tends to $P(A)$ by the monotone convergence theorem.

"$6 \implies 3$" Let $f : E \to \mathbb{R}$ be a bounded measurable function and $\mathcal{U}_f$ the set of points where it is not continuous. We only consider the case where $P(\mathcal{U}_f) = 0$.

Let $\varepsilon > 0$. We want to prove that there exist finitely many values $y_1, \ldots, y_m \in \mathbb{R}$ and corresponding disjoint sets $E_1, \ldots, E_k \subset E$, $E_i \cap E_j = \emptyset$ for $i \neq j$, such that the step function $g := \sum_{k=1}^m y_k \mathbb{1}_{E_k}$ is a good approximation to the measurable function, in the sense that $\|f - g\|_\infty < \varepsilon$. Additionally, we require that the measure of the boundaries of the sets vanishes, $P(\partial E_k) = 0$.

If this can be shown, then we can estimate the integral by the triangle inequality

$$\begin{aligned}\left|\int f \, dP - \int f \, dP_n\right| &\leq \left|\int (f - g) \, dP\right| + \left|\int g \, dP - \int g dP_n\right| + \left|\int (f - g) \, dP_n\right| \\ &\leq \varepsilon + \sum_{k=1}^m |y_m| \, |P(E_k) - P_n(E_k)| + \varepsilon.\end{aligned}$$

Since the middle sum is finite and $P(\partial E_k) = 0$, we can apply the assumption and see that the term in the middle becomes smaller than ε if the index n is large enough.

To obtain the values $y_1, \ldots, y_m$, note that any covering of the interval $[-\|f\|_\infty, \|f\|_\infty] \subset \mathbb{R}$ by disjoint intervals $[y_k, y_{k+1})$ of width $|y_k - y_{k+1}| < \varepsilon$ will provide a good approximation if we set $E_k := f^{-1}([y_k, y_{k+1}))$. However, we have to ensure that $P(\partial E_k) = 0$, which requires a little more work.

First, note that for any metric space E, we have

$$\partial f^{-1}(B) \subset f^{-1}(\partial B) \cup \mathcal{U}_f.$$

In particular, we have

$$P(\partial E_k) \leq P\left(f^{-1}(\partial [y_k, y_{k+1})\right) + P(\mathcal{U}_f) = P\left(f^{-1}(y_k)\right) + P\left(f^{-1}(y_{k+1})\right) + P(\mathcal{U}_f).$$

The last term vanishes by assumption. Now, if we can choose the points y_k such that the preimage has vanishing measure, $P(f^{-1}(y_k)) = 0$, then we are done.

To do this, observe that the set of values $C_n = \{y \in \mathbb{R} : P\left(f^{-1}(y)\right) > 1/n\}$ is finite, because the preimages $f^{-1}(y)$ are disjoint and their total measure may not exceed $P(E) = 1$. Hence,

the union $C = \bigcup_{n=1}^{\infty} C_n = \{y \in \mathbb{R} : P\left(f^{-1}(y)\right) > 0\}$ is at most countably infinite. Since there are uncountably many choices available for the points y_k, we can certainly arrange that $y \in \mathbb{R} \setminus C$ as desired. □

In the special case where the underlying space is the line of real numbers, $E = \mathbb{R}$, we can also characterize the weak convergence in terms of the cumulative distribution function.

Definition. A sequence $(F_n)_{n=1}^{\infty}$ of cumulative distribution functions **converges weakly** to a cumulative distribution function F if and only if it converges at every point where the limit is continuous,

$$\boxed{F_n \Rightarrow F :\iff F_n(x) \to F(x) \text{ for all } x \in \mathbb{R} \text{ where } F \text{ is continuous.}}$$

Proposition 3.3 (**Weak convergence** and **cumulative distribution function**). *A sequence of probability measures $(P_n)_{n=1}^{\infty}$ on the real line converges weakly to a probability measure P if and only if the corresponding sequence of distribution functions $(F_n)_{n=1}^{\infty}$ converges weakly to a distribution function F,*

$$\boxed{P_n \Rightarrow P \quad \iff \quad F_n \Rightarrow F.}$$

Proof. "$\Longrightarrow$" Let $\varepsilon > 0$ and let x be a real number. With the Portemanteau theorem 3.2, we see that

$$\begin{aligned} F(x-\varepsilon) = P((-\infty, x-\varepsilon]) &\leq P((-\infty, x)) \leq \liminf_{n\to\infty} P_n((-\infty, x)) \leq \liminf_{n\to\infty} P_n((-\infty, x]) \\ &\leq \limsup_{n\to\infty} P_n((-\infty, x]) \leq P((-\infty, x]) = F(x). \end{aligned}$$

If the function F is continuous at the point x, then the limes inferior and the limes superior agree, and we have $F_n(x) \to F(x)$.

"$\Longleftarrow$" We use the Portemanteau theorem again and show that the integrals $\int f\, dP_n$ converge for every bounded Lipschitz-continuous function f.

Let $\varepsilon > 0$. The cumulative distribution F is continuous at a point x if and only if $P(\{x\}) = 0$. Since $P(\mathbb{R}) = 1$ is finite, this means that there are at most countably many points where the cumulative distribution is not continuous. Hence, we can find points $y_1 < y_2 < \ldots < y_N$ such that F is continuous at these points, $F(y_N) - F(y_1) \geq 1 - \varepsilon$ and $|y_{k+1} - y_k| < \varepsilon$ for every index k. Consider the step function $g(x) = \sum_{k=1}^{N-1} f(y_k) \mathbb{1}_{(y_k, y_{k+1}]}(x)$. By assumption, we have

$$\int g\, dP_n = \sum_{k=1}^{N-1} f(y_k)(F_n(y_{k+1}) - F_n(y_k)) \to \sum_{k=1}^{N-1} f(y_k)(F(y_{k+1}) - F(y_k)) = \int g\, dP.$$

Since the function f was assumed to be Lipschitz-continuous with Lipschitz constant L, the step function is a good approximation in the sense that

$$\int |f - g|\, dP_n \leq L\varepsilon + 2\|f\|_{\infty}[1 - (F_n(y_N) - F_n(y_1))].$$

If we choose the index n large enough, then we can make this smaller than a constant times ε. Then, the desired convergence follows from the triangle inequality. □

We now want to consider the question of compactness: When does an arbitrary family of probability distributions have a subsequence that converges in the weak sense?

Definition. A family of probability measures $\mathcal{P}$ on a metric space is called **relatively sequentially weakly compact** if every sequence $(P_n)_{n=1}^{\infty}$ of measures from this family, $P_n \in \mathcal{P}$, has a weakly convergent subsequence, $P_{n_k} \Rightarrow P$ for $k \to \infty$.

("Relative" refers to the fact that the limit P does not have to be a member of the family again. "Sequentially" refers to the consideration of sequences and subsequences. "Weakly" indicates that the notion of convergence considered here is that of weak convergence.)

It turns out that only a small condition of uniformity is needed:

Definition. A family of probability measures $\mathcal{P}$ on a metric space is called **tight** if, for all $\varepsilon > 0$, there exists a compact set $K \subset E$ such that $P(K^c) < \varepsilon$ for all members of the family, $P \in \mathcal{P}$.

Lemma 3.4 (**Tightness** of **finite families**). *On a polish space, every finite family of probability distributions is tight.*

Proof. Since a finite union of compact sets is compact, it is enough to show that any family consisting of a single probability distribution P is tight. Let $\varepsilon > 0$.

Since the polish space E is separable, we can find, for each $n \in \mathbb{N}$, a sequence of points $x_1^n, x_2^n, \ldots$ such that the space is the union $E = \bigcup_{k=1}^{\infty} B_{1/n}(x_k^n)$, where $B_{1/n}(x)$ denotes an open ball of radius $1/n$ around the point x. Since these balls exhaust the space, we can find a number N_n such that the union of the first N_n balls, $A_n = \bigcup_{k=1}^{N_n} B_{1/n}(x_k^n)$, satisfies

$$P\left(A_n^c\right) < \varepsilon 2^{-n}.$$

By construction, the set $A = \bigcap_{n=1}^{\infty} A_n$ is totally bounded. Hence, its closure $\overline{A}$ is compact, but its complement satisfies

$$P\left(\overline{A}^c\right) \leq P\left(A^c\right) \leq \sum_{n=1}^{\infty} P\left(A_n^c\right) \leq \varepsilon$$

as desired. □

Theorem 3.5 (**Prokhorov's theorem**). *Let E be a polish space. Then, a family $\mathcal{P}$ of probability measures on this space is relatively sequentially weakly compact if and only if it is tight.*

The main difficulty in Prokhorov's theorem is to show that tightness implies the existence of a weakly converging subsequence. We will not prove it in full generality here, but we will prove it for the special case $E = \mathbb{R}$. In this case, we can use the cumulative distribution function and the following proposition:

Proposition 3.6 (**Helly's selection theorem**). *Let $(F_n)_{n=1}^{\infty}$ be a sequence of distribution functions. Then, there exists a subsequence $(F_{n_k})_{k=1}^{\infty}$ and a monotonically increasing, right-continuous function $F : \mathbb{R} \to [0, 1]$ such that $F_{n_k}(x) \to F(x)$ at every point x where the function F is continuous.*

Remark. Note that this limit F does *not* have to be the cumulative distribution of a probability measure! It may well be that the total weight is smaller than one, $F(\infty) - F(-\infty) < 1$. For instance, if $F_n = G(x - n)$ for some distribution function $G(x)$, then the pointwise limit is a constant, $F_n \to 0$.

Proof. We use a diagonal argument to construct a candidate for the limit. Let $q_1, q_2, \ldots$ be an enumeration of the rational numbers $\mathbb{Q} \subset \mathbb{R}$. Since the sequence of values $F_n(q_1)$ is bounded, there exists a subsequence that converges at this point, $F_{n_1(k)}(q_1) \to f_1 \in [0, 1]$. In turn, this subsequence has a subsequence such that converges at the second point, $F_{n_2(k)}(q_2) \to f_2$. Repeat this process for every point and consider the diagonal sequence $F_{n(k)} = F_{n_k(k)}$. It has a limit $F_{n(k)}(q) \to f(q)$ for every rational number $q \in \mathbb{Q}$.

Since any member of the sequence satisfies $F_{n(k)}(q) \leq F_{n(k)}(q')$ for any pair of rational numbers $q \leq q'$, we also have $f(q) \leq f(q')$ in the limit. Hence we can extend the limit to a monotonically increasing function that is defined for all real numbers,

$$F : \mathbb{R} \to [0, 1], \quad F(x) = \inf\{f(q) : q \geq x, q \text{ rational}\}.$$

It satisfies $F(q) = f(q) = \lim_{k\to\infty} F_{n(k)}(q)$ for every rational point $q \in \mathbb{Q}$. It is straightforward to check that this function is right-continuous.

It remains to show that the sequence converges $F_n(x) \to F(x)$ at every point where the function F is continuous. But this follows from the fact that for each $\varepsilon > 0$, we can find rational numbers q_1, q_2 with $q_1 < x < q_2$ such that both $F(x) - F(q_1) \leq \varepsilon$ and $F(q_2) - F(x) \leq \varepsilon$. By construction, we know that $F_{n(k)}(q_{1,2}) \to F(q_{1,2})$ and the convergence $F_{n(k)}(x) \to F(x)$ follows from the triangle inequality. □

Proof of Prokhorov's theorem (partial). "$\Longrightarrow$" To show tightness, we will construct compact sets by using an argument similar to the one we used in the proof of Lemma 3.4, where we showed the tightness of any finite family.

Let $x_1, x_2, \ldots$ be a sequence of points that are dense in the polish space E. Consider the union of open balls $A_{n,N} = \bigcup_{k=1}^{N} B_{1/n}(x_k)$. Since the points are dense, these sets exhaust the space, $E = \bigcup_{N=1}^{\infty} A_{n,N}$ for each fixed size n.

Let $\varepsilon > 0$. If we can show that

$$\forall n, \exists N_n, \forall P \in \mathcal{P},\ P(A_{n,N}^c) \leq \varepsilon 2^{-n}, \tag{2}$$

then we can consider the set $A = \bigcap_{n=1}^{\infty} A_{n,N_n}$. This set is totally bounded, hence its closure $\overline{A}$ is compact. By the previous assertion, for each probability measure in the family, $P \in \mathcal{P}$, we have

$$P(\overline{A}^c) \leq P(A^c) \leq \sum_{n=1}^{\infty} P(A_{n,N_n}^c) \leq \varepsilon,$$

which means that the family is tight.

To prove the assertion (2), assume the contrary. This means that there would be some $n > 0$ such that we could find a sequence of probability measures $(P_N)_{N=1}^{\infty}$ from the family with $P_N(A_{n,N}^c) > \varepsilon 2^{-n} =: \delta$. By assumption, this sequence converges weakly to some probability measure, $P_N \Rightarrow P$.

Since the complements are contained in each other, $A^c_{n,1} \supset A^c_{n,2} \supset \dots$, we have the inequality $P_N(A^c_{n,M}) \geq P_N(A^c_{n,N}) > \delta$ whenever $N \geq M$. Moreover, the sets are closed. By the Poremanteau theorem, Proposition 3.2, we conclude that

$$\delta \leq \limsup_{N\to\infty} P_N(A^c_{n,M}) \leq P(A^c_{n,M})$$

for every fixed index M. But the sequence of sets $A_{n,M}$ exhausts the space E, and the right-hand side must tend to zero in the limit $M \to \infty$, a contradiction.

"$\Longleftarrow$" We only prove this for the special case $E = \mathbb{R}$. Let $(P_n)_{n=1}^\infty$ be a sequence of probability measures that is tight. Consider the corresponding cumulative distribution functions $(F_n)_{n=1}^\infty$. By Helly's selection theorem, a subsequence to a function F, where $F_{n_j}(x) \to F(x)$ at every point x where the function F is continuous. We only have to show that this limit is the cumulative distribution of a probability measure, which is tantamount to showing that $F(\infty) - F(-\infty) = 1$.

But since the family is tight, for each $\varepsilon > 0$, we can find a point $x \in \mathbb{R}$ such that F is continuous at x and

$$P_n([-x, x]) = F_n(x) - F_n(-x) \geq 1 - \varepsilon$$

for each index n. But this is also true in the limit. Letting $\varepsilon \to 0$, we see that $F(\infty) - F(-\infty) = 1$ as desired. □

In the definition of weak convergence, one has to consider the integrals over *all* continuous functions. It would be beneficial if we could restrict our attention to subsets of functions. The minimum requirement is the following:

Definition. Let E be a polish space. A family of bounded continuous functions $\mathcal{C}$ is called a **separating family** if any probability distribution is uniquely determined by their integrals, that is

$$\boxed{\int f\, dP_1 = \int f\, dP_2 \text{ for all } f \in \mathcal{C} \quad \Longrightarrow \quad P_1 = P_2}$$

for any two probability distributions P_1, P_2.

Example. The collection of all bounded continuous functions is a separating family. This is equivalent to the uniqueness of the weak limit, Lemma 3.1.

Example. The collection of all bounded Lipschitz-continuous functions is a separating family. This follows e.g. from the Portemanteau theorem.

It turns out that together with tightness, this requirement is already sufficient:

Proposition 3.7 (**Weak convergence** from **separating family**)**.** *Let $(P_n)_{n=1}^\infty$ be a sequence of probability distributions on a polish space. Then, this sequence converges weakly to a probability distribution, $P_n \Rightarrow P$, if and only if the sequence is tight and there exists a separating family $\mathcal{C}$ such that the integrals converge, $\int f\, dP_n \to \int f\, dP$, for all members the family $f \in \mathcal{C}$.*

Proof. "$\Longleftarrow$" By Prokhorov's theorem, a weakly converging sequence is tight. Convergence of the integrals is also clear.

"$\Longrightarrow$" The main idea is to show that any subsequence has a weakly converging subsequence, and all limits obtained in this way are equal.

Let g be any bounded continuous function. The sequence of real numbers $\int g\,dP_n$ is bounded by $\|g\|_\infty$ and hence has at least one limit point. We want to show that the real number $\int g\,dP$ is, in fact, the only limit point. To that end, let $b \in \mathbb{R}$ be any limit point of this sequence, so that $\int g\,dP_{n_k} \to b$ for some subsequence P_{n_k}. This subsequence is also tight, and we can apply Prokhorov's theorem to conclude that a subsequence of this subsequence converges weakly, $P_{n_{k_j}} \Rightarrow Q$. By assumption, we have $\int f\,dP = \int f\,dQ$ for all functions in the separating family $\mathcal{C}$, and it follows that $Q = P$. In other words, we have $\int g\,dP_{n_{k_j}} \to \int g\,dP = b$. □

3.2 Characteristic functions

Characteristic functions are an extremely useful tool for calculating with probability distributions. They correspond to the Fourier transform of probability measures.

Definition. Let P be a probability distribution on the n-dimensional euclidean space $(\mathbb{R}^n, \mathcal{B})$. Then, its **characteristic function** or **Fourier-transform** is defined as the function

$$\boxed{\hat{P} : \mathbb{R}^n \to \mathbb{C}, \quad \hat{P}(t) = E\big[e^{it\cdot x}\big] = \int_{\mathbb{R}^n} e^{it\cdot x}\,dP(x).}$$

Definition. Let $X : \Omega \to \mathbb{R}^n$ be a vector-valued random variable. Then, its **characteristic function** ϕ is the characteristic function of the corresponding probability distribution induced on the space $(\mathbb{R}^n, \mathcal{B})$,

$$\boxed{\phi_X : \mathbb{R}^n \to \mathbb{C}, \quad \phi_X(t) = E\big[e^{it\cdot X}\big]\,.}$$

Notation. We often denote the characteristic function with ϕ. Note that every probability distribution on the euclidean space is the probability distribution of a random variable.

Remark. The characteristic function of a probability distribution always satisfies

$$|\phi(t)| \leq 1 \text{ for all } t \in \mathbb{R}^n \quad \text{ and } \phi(0) = 1.$$

The Fourier-transform has some very nice algebraic properties, in particular when it comes to the convolution of two probability distributions, which corresponds to the sum of independent random variables:

Lemma 3.8 (**Algebraic properties** of **characteristic functions**). *The characteristic functions of random variables have the following algebraic properties:*

1. *ϕ_X real-valued $\iff$ X and $-X$ have the same probability distribution.*
2. *$\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.*
3. *If X and Y are independent random variables, then the characteristic function of the sum is given by the product, $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.*

Proof. On 3. The independence of the variables X and Y implies that the random variables e^{itX} and e^{itY} are independent. It follows that $E\big[e^{it(X+Y)}\big] = E\big[e^{itX}e^{itY}\big] = E\big[e^{itX}\big]\,E\big[e^{itY}\big]$. □

Example. The standard normal distribution has the probability density $p(x)dx = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$. Its characteristic function can be calculated as follows:

$$\begin{aligned}\phi_{\mathcal{N}(0,1)}(t) &= \int_{-\infty}^{\infty} e^{itx}p(x)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-it)^2}{2}-\frac{t^2}{2}}dx \\ &= e^{-\frac{t^2}{2}}.\end{aligned}$$

In the last line, we have shifted the integration contour in the complex plane by $-it$ and used the fact that the complex function $e^{-\frac{z^2}{2}}$ is holomorphic. Using the algebraic properties, we obtain the characteristic function of a normal distribution with expectation μ and variance σ^2:

$$\phi_{\mathcal{N}(\mu,\sigma^2)} = e^{it\mu-\frac{t^2\sigma^2}{2}}.$$

Another important property of the Fourier transform is that it converts bounds on the random variable into regularity of the characteristic function:

Lemma 3.9 (**Regularity** of **characteristic functions**). *The characteristic function ϕ of a random variable X has the following regularity properties:*

1. *The characteristic function is uniformly continunous.*
2. *If the random variable has a finite k-th moment, $E\big[|X|^k\big] < \infty$, then the probability distribution is k-times continously differentiable with*

   $$\partial_t^\alpha \phi_X(t) = E\big[(iX)^\alpha e^{itX}\big] \qquad \text{for } |\alpha| \le k.$$

   (Here, $\alpha = (\alpha_1, \ldots, \alpha_n)$ is written using multi-index notation.)

   If the random variable is real-valued, then it has a Taylor series

   $$\phi_X(t) = 1 + E[X]\,it + \frac{1}{2!}E\big[X^2\big]\,(it)^2 + \ldots + \frac{i^k}{k!}E\Big[X^k\Big]\,t^k + o(t^k).$$

3. *If the moments of a real-valued random variable satisfy*

   $$\limsup_{n\to\infty}\left(\frac{1}{n}\sqrt[n]{E[|X|^n]}\right) < \infty,$$

   then the characteristic function is real-analytic.

   A sufficient condition for this inequality is that $E\big[e^{s|X|}\big] < \infty$ for one value $s > 0$.

Proof. ▷ todo ◁ □

The characteristic function of a probability distribution uniquely determines that distribution. This follows from the following lemma:

Lemma 3.10 (**Trigonometric functions** are a **separating family**). *The collection of functions*

$$\mathcal{C} = \{f \in C_b(\mathbb{R}^n) : f(x) = e^{ik\cdot x}, k \in \mathbb{R}^n\}$$

is a separating family on the polish space $\mathbb{R}^n$.

Proof. Let P_1, P_2 be two probability measures. It is enough to show that if all integrals $\int f\,dP_1$ and $\int f\,dP_2$ are equal for trigonometric functions f, then they are equal for any bounded continuous function f.

Let f be a bounded continuous function and $\varepsilon > 0$. Since any finite family of probability measures on a polish space is tight, we can find a compact set $K \subset \mathbb{R}^n$ such that the integrals outside this set satisfy $\int_{K^c} |f|\,dP_j \leq \varepsilon\|f\|_\infty$ for $j = 1, 2$. By the Weierstrass approximation theorem for continuous functions on compact sets, we can approximate the restriction $f|_K$ by a trigonometric polynomial, i.e. a linear combination $g \in \operatorname{span}\{e^{ikx} : k \in \mathbb{R}^n\}$ with the property $\|f|_K - g|_K\|_\infty < \varepsilon$. Then, we can use the triangle inequality to estimate

$$\begin{aligned}\left|\int_K f\,dP_1 - \int_K f\,dP_2\right| &\leq \left|\int_K f\,dP_1 - \int_K g\,dP_1\right| + \left|\int_K g\,dP_1 - \int_K g\,dP_2\right| + \left|\int_K g\,dP_2 - \int_K f\,dP_2\right| \\ &\leq \varepsilon + 0 + \varepsilon = 2\varepsilon \\ \left|\int f\,dP_1 - \int f\,dP_2\right| &\leq \left|\int_K f\,dP_1 - \int_K f\,dP_1\right| + \left|\int_{K^c} f\,dP_1 - \int_{K^c} f\,dP_1\right| \\ &\leq 2\varepsilon + 2\varepsilon\|f\|_\infty.\end{aligned}$$

Since ε was arbitrary, the desired conclusion follows. □

Corollary. Two probability measures on $\mathbb{R}^n$ are equal if and only if their characteristic functions are equal.

If we consider a sequence of probability distributions, then the convergence of their characteristic functions tells us a lot about the weak convergence of this sequence:

Proposition 3.11 (**Convergence of characteristic functions** and **weak convergence**). *Let* $(P_n)_{n=1}^\infty$ *be a sequence of probability measures on* $\mathbb{R}^n$ *with characteristic functions* $(\phi_n(t))_{n=1}^\infty$.

1. *If the probability measures converge weakly,* $P_n \Rightarrow P$, *then the characteristic functions converge locally uniformly* $\phi_n(t) \to \phi(t)$.
2. *If the characteristic functions converge* pointwise everywhere $\phi_n(t) \to \phi(t)$, *and the limit* $\phi(t)$ *is continuous at* $t = 0$, *then there exists a probability measure* P *such that the probability measures converge weakly,* $P_n \Rightarrow P$.

In both cases, the limit $\phi(t)$ *is the characteristic function of the probability measure* P.

Proof. For simplicity, we only consider the case of probability distributions on the space $\mathbb{R}$. The higher-dimensional case is completely analogous.

On 1. From the definition of weak convergence, it is clear that the characteristic functions converge pointwise. To show locally uniform convergence, let $\varepsilon > 0$ and consider

$$|\phi_n(t) - \phi_n(s)| = \left|\int e^{isx}(e^{i(t-s)x} - 1)\,dP_n(x)\right| \leq \int \left|e^{i(t-s)x} - 1\right| dP_n(x).$$

Since the sequence of probability distributions is right, there is a radius R such that $P_n(B_R(0)^c) < \varepsilon$ for all indices n. If we choose $|t-s| < \varepsilon/R$ and the point x inside the ball, then the exponent satisfies $|(t-s)x| < \varepsilon$ and we can conclude

$$|\phi_n(t) - \phi_n(s)| \leq \int_{|x|\leq R} \left| e^{i(t-s)x} - 1 \right| \, dP_n(x) + 2\varepsilon \leq \int_{|X|\leq R} \varepsilon \, dP_n(x) + 2\varepsilon \leq 3\varepsilon.$$

This estimate is uniform in n and an application of the triangle inequality establishes locally uniform convergence.

On 2. To establish weak convergence, we want to use Proposition 3.7, a consequence of Prokhorov's theorem. We have already seen that the collection of trigonometric functions is a separating family, so we only have to establish tightness. We show that this follows from the continuity of the limit $\phi(t)$ at $t = 0$.

Consider the family of functions

$$h_\varepsilon(t) := \frac{1}{\varepsilon} h\left(\frac{t}{\varepsilon}\right) \quad \text{where} \quad h(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad \text{and } \varepsilon > 0.$$

This is a *δ-family*, which means that

$$\int h_\varepsilon(t) \, dt = 1 \quad \text{and} \quad \lim_{\varepsilon \to 0} \int h_\varepsilon(t)\phi(t) \, dt = \phi(0)$$

for any bounded function ϕ that is continous at $t = 0$. By the Lebesgue dominated convergence theorem, we can conclude

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \int h_\varepsilon(t)\phi_n(t) \, dt = \phi(0) = 1 \tag{3}$$

for the characteristic functions.

However, using Fubini's theorem, we can express the integral in terms of the reverse Fourier transform of the function $h_\varepsilon(t)$, which we have already calculated:

$$\begin{aligned} \int h_\varepsilon(t)\phi_n(t) \, dt &= \int h_\varepsilon(t) e^{itx} \, dt \, dP_n(x) = \frac{1}{\sqrt{2\pi\varepsilon^2}} \int e^{-\frac{t^2}{2\varepsilon^2}} e^{itx} \, dt \, dP_n(x) \\ &= \int e^{-\frac{(x\varepsilon)^2}{2}} dP_n(x). \end{aligned}$$

Now, if the family of probability distributions P_n were not tight, then there would exist an $\alpha > 0$ so that for every radius $R > 0$, we would have $P_n([-R, R]^c) > \alpha$ for at least one index n. Since every finite family is tight, this must actually hold for infinitely many indices n, otherwise we could enlarge the radius. But given $\varepsilon > 0$, we can choose the product $R\varepsilon$ so large that $e^{-\frac{(R\varepsilon)^2}{2}} < 1/2$, and obtain that

$$1 - \int_{\mathbb{R}} e^{-\frac{(x\varepsilon)^2}{2}} dP_n(x) = \int_{\mathbb{R}} (1 - e^{-\frac{(x\varepsilon)^2}{2}}) dP_n(x) \geq \int_{[-R,R]^c} (1 - e^{-\frac{(x\varepsilon)^2}{2}}) dP_n(x) > \frac{\alpha}{2}$$

for infinitely many indices n. This is a contradiction to the limit equation (3).

To put it differently, the point of the proof is that the factor $e^{-\frac{(x\varepsilon)^2}{2}}$ decays for large x, regardless of how small ε is. □

Example. The condition that the limit $\phi(t)$ be continuous at $t = 0$ is important and cannot be dropped. As an example, let us consider the sequence $(P_n)_{n=1}^{\infty}$ of uniform probability distribution the interval $[-n, n]$, that is $P_n = \frac{1}{2n}\mathbb{1}_{[-n,n]}dx$. The corresponding characteristic functions are

$$\phi_n(t) = \frac{1}{2n}\int_{-n}^{n} e^{itx}\,dx = \frac{\sin(nt)}{nt}.$$

For $n \to \infty$, they converge to the function

$$\phi(t) = \begin{cases} 1, & \text{if } t = 0 \\ 0, & \text{otherwise} \end{cases}.$$

The limit is not continuous at $t = 0$, and indeed, this sequence of probability distributions does not converge weakly, because it is not tight.

3.3 Lévy's Theorem

With the new tool of characteristic functions at our disposal, we go back to series of independent random variables, Section 2.7. Besides Kolmogorov's maximal inequality, that ultimately allowed us to prove the strong law of large numbers, we also have the following important characterization of convergence of a series:

Theorem 3.12 (**Lévy's Theorem**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of* independent *random variables and consider the partial sums $S_n = X_1 + X_2 + \ldots + X_n$. Then, the following are equivalent:*

1. *The probability distributions α_n of the partial sums S_n converge weakly to a probability distribution α on $\mathbb{R}$.*
2. *The partial sums S_n converge in probability to a random variable $S(\omega)$.*
3. *The partial sums S_n converge almost surely to a random variable $S(\omega)$.*

In other words, in this case, weak notions of convergence already imply strong notions of convergence. The main technical tool for establishing these results is

Lemma 3.13 (**Lévy's Inequality**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of* independent *random variables with no assumptions on integrability and $S_n = X_1 + X_2 + \ldots + X_n$ the corresponding partial sums. Then, we have the implication*

$$\boxed{\sup_{0\le k\le n} P\left(|S_n - S_k| > \frac{l}{2}\right) \le \delta \quad \Longrightarrow \quad P\left(\sup_{1\le k\le n} |S_k| \ge l\right) > \frac{\delta}{1-\delta}.} \tag{4}$$

Proof. As in the proof of Kolmogorov's maximal inequality 2.14, consider the events

$$E_k := \{|S_1| \le l, |S_2| \le l, \ldots, |S_k| > l\}, \quad k = 1 \ldots n$$

where the first sums are smaller than the bound l, but the k-th sum exceeds it. These events are mutually disjoint, and we have

$$E := \left\{\sup_{1\le k\le n} |S_k(\omega)| > l\right\} = \biguplus_{k=1}^{n} E_k.$$

On one hand, we may use independence to reason

$$\begin{aligned} P\left(E \cap |S_n| \leq \frac{l}{2}\right) &= \sum_{k=1}^{n} P\left(E_k \cap |S_n| \leq \frac{l}{2}\right) \\ &\leq \sum_{k=1}^{n} P\left(E_k \cap |S_n - S_k| > \frac{l}{2}\right) = \sum_{k=1}^{n} P(E_k) P\left(|S_n - S_k| > \frac{l}{2}\right) \\ &\leq P(E)\delta. \end{aligned}$$

On the other, the assumptions also imply

$$P(E) = P\left(E \cap |S_n| \leq \frac{l}{2}\right) + P\left(E \cap |S_n| > \frac{l}{2}\right) \leq P\left(E \cap |S_n| \leq \frac{l}{2}\right) + \delta.$$

Taking both together yields

$$P(E) \leq \delta P(E) + \delta \implies P(E) \leq \frac{\delta}{1-\delta}$$

as desired. □

Proof of Lévy's theorem. It is clear that the strong notions of convergence imply the weak ones. We have to establish the other directions.

"1 $\implies$ 2" We use characteristic functions. Let $\phi_k(t)$ denote the characteristic function of the random variable X_k. Since the summands are independent, the characteristic function of the partial sum S_n is given by the product $\prod_{k=1}^{n} \phi_k(t)$, and the characteristic function of the remainder $S_n - S_m = X_{m+1} + X_{m+2} + \ldots + X_n$ is given by the product $\prod_{k=m}^{n} \phi_k(t)$.

By assumption, the first product converges locally uniformly to the characteristic function of the probability distribution α:

$$\lim_{n\to\infty} \prod_{k=1}^{n} \phi_k(t) = \phi(t).$$

The limit is continuous with $\phi(0) = 1$. Hence, there exists an open neighborhood U of $t = 0$ such that $\phi(t) > 1/2$ in this neighborhood. Since the characteristic functions have magnitude at most one, $|\phi_k(t)| \leq 1$, each partial product must satisfy $|\prod_{k=1}^{n} \phi_k(t)| > 1/2$ in this neighborhood $t \in U$, so that the limit can have a large enough magnitude. This proves that the implication

$$\left|\prod_{k=1}^{n} \phi_k(t) - \prod_{k=1}^{m-1} \phi_k(t)\right| < \varepsilon \implies \left|1 - \prod_{k=m}^{n} \phi_k(t)\right| < 2\varepsilon \quad \text{for all } t \in U$$

holds for any index pair n, m with $n \geq m$.

Of course, the left-hand side of this implication comes from Cauchy criterion for the existence of the limit. Applying Lemma 3.14 below repeatedly proves that we can upgrade the right-hand side to a Cauchy criterion that holds for *all* $t \in \mathbb{R}$. More precisely, for each point $t \in \mathbb{R}$ and for all $\varepsilon > 0$, we can find an N such that for all indices $n, m \geq N$, we have $|1 - \prod_{k=m}^{n} \phi_k(t)| < \varepsilon$. This means that for each sequence of index pairs (m_j, n_j) with $m_j \to \infty$, the characteristic

function of the random variables $(S_{n_j} - S_{m_j})$ converges to constant one. By Proposition 3.11 on the pointwise convergence of characteristic functions, we conclude that this random variable converges to zero *in probability*.

In the end, we want to prove the Cauchy criterion in probability and use Lemma 2.10 to establish convergence in probability. Assume that the Cauchy criterion in probability does not hold. This means that there exist $\varepsilon, \delta > 0$ such that we can find a sequence of index pairs (m_j, n_j) with $m_j \to \infty$ which satisfy $P\left(\left|S_{n_j} - S_{m_j}\right| > \varepsilon\right) > \delta$. But this is a contradiction to the previously established fact that these differences converge to zero in probability.

"2 $\implies$ 3" This is essentially a direct consequence of Lévy's inequality. By assumption, the partial sums S_n are a Cauchy sequence in probability, that is for all $\varepsilon, \delta > 0$, there exists an index N such that

$$\forall n, m \geq N,\ P\left(|S_n - S_m| \geq \varepsilon\right) \leq \delta.$$

If $M \geq N$ is some other index, then this certainly holds for all $n, m \geq M$. We can write $S_n - S_m = (S_n - S_M) - (S_m - S_M)$ and apply Lévy's inequality to upgrade this to

$$\forall \varepsilon, \delta > 0,\ \exists N, \forall n, M \geq N,\ P\left(\sup_{M \leq k \leq n} |S_k - S_M| \geq 2\varepsilon\right) \leq \frac{\delta}{1-\delta}.$$

By Lemma 2.12, a consequence of the Borel-Cantelli lemma, this implies convergence almost surely. □

Lemma 3.14 (A **cosine estimate**)**.** *For all $t \in \mathbb{R}$, we have*

$$1 - \cos(2t) \leq 4[1 - \cos(t)].$$

Corollary 3.15. *If $\phi(t)$ is the characteristic function of a probability distribution, we have*

$$1 - \operatorname{Re}\phi(2t) \leq 4[1 - \operatorname{Re}\phi(t)].$$

Proof. If we expand the cosine to second order in t, so that $\cos t = 1 - \frac{1}{2}t^2 + o(t^2)$, the inequality becomes $4\frac{1}{2}t^2 \leq 3 + 4\frac{1}{2}t^2$, which is true for small arguments. For the general case, we use the double angle formula $\cos(2t) = 2\cos^2 t - 1$ and calculate

$$\begin{aligned} 1 - \cos(2t) = 2 - 2\cos^2 t &\overset{?}{\leq} 4(1 - \cos t) \\ 0 &\overset{?}{\leq} \cos^2 t - 2\cos t + 1 = (\cos t - 1)^2. \end{aligned}$$

Since the right-hand side is a sum of squares, the inequality holds. □

3.4 The Central Limit Theorem

In the previous chapter, we have studied the law of large numbers, which states that under very general conditions, the average of a sequence of i.i.d. random variables $(X_n)_{n=1}^{\infty}$ tends to the expectation value,

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to E[X_n],$$

in probability or even almost surely.

We now want to understand the asymptotics of this convergence. If we denote the variance of the random variables by $V[X_k] = \sigma^2$, then a very simple calculation shows that

$$V\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n^2}\sum_{k=1}^{n} V[X_k] = \frac{1}{n}\sigma^2,$$

so the standard deviation of the average from its mean μ goes like the inverse square root, $\frac{1}{\sqrt{n}}\sigma$. But it turns out that the asymptotic distribution of the average to order $1/\sqrt{n}$ can be determined exactly. This is the content of the celebrated central limit theorem:

Theorem 3.16 (**Central Limit Theorem**). *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables in L^2 with $E[X_n] = \mu$ and $V[X_n] = \sigma^2$. Furthermore, let*

$$S_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} \frac{X_k - \mu}{\sigma}$$

be the sequence of normalized partial sums, that is $E[S_n] = 0$ and $V[S_n] = 1$. Then, their probability distribution converges weakly to the standard normal distribution:

$$\boxed{\lim_{n\to\infty} P\left(a \le S_n \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\, dx.} \tag{5}$$

Proof. We use characteristic functions. Let $\phi(t)$ be the characteristic function of the random variable $Y_n = (X_n - \mu)/\sigma$. Then, $\phi(t/\sqrt{n})^n$ is the characteristic function of the normalized partial sum S_n, and we want to prove that this converges to the characteristic function of the standard normal distribution,

$$\phi\left(\frac{t}{\sqrt{n}}\right)^n \to e^{-\frac{t^2}{2}} \qquad \text{for every } t \in \mathbb{R}.$$

Since the random variables are in L^2, the characteristic function has regularity C^2 and thus has a Taylor expansion

$$\phi(t) = 1 + itE[Y_n] - \frac{1}{2}t^2 E[Y_n^2] + o(t^2) = 1 - \frac{1}{2}t^2 + o(t^2) \qquad \text{for } t \to 0.$$

Using the Taylor expansion of the logarithm,

$$-\ln(1-x) = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \ldots = x + o(x) \qquad \text{for } x \to 0,$$

we obtain the following asymptotics for $n \to \infty$:

$$\begin{aligned}\phi\left(\frac{t}{\sqrt{n}}\right)^n &= \exp\left(n \ln \phi\left(\frac{t}{\sqrt{n}}\right)\right) = \exp\left(n \ln\left[1 - \frac{1}{2n}t^2 + o(1/n)\right]\right) \\ &= \exp\left(-\frac{1}{2}t^2 + o(1)\right).\end{aligned}$$

This converges to the characteristic function of standard normal distribution, as desired. □

In the central limit theorem, it is not strictly necessary that the random variables are identically distributed or have the same variance. If we consider suitably normalized partial sums, we may still conclude that the central limit theorem holds under reasonable conditions:

Proposition 3.17 (**Lindeberg's theorem**). *Let $(X_k)_{k=1}^{\infty}$ be a sequence of independent random variables with mean zero, $E[X_k] = 0$ and variance $\sigma_k^2 = V[X_k]$. Let α_k denote its distribution. Furthermore, let $s_n^2 = \sum_{k=1}^{n} \sigma_k^2$ be the sum of variances and*

$$S_n = \frac{1}{s_n} \sum_{k=1}^{n} X_n$$

*be the normalized partial sum. If we assume that $s_n \to \infty$ for $n \to \infty$ and if **Lindeberg's condition***

$$\boxed{\text{for all } \varepsilon > 0, \quad \lim_{n\to 0} \sum_{k=1}^{n} \frac{1}{s_n^2} \int_{|x| \geq \varepsilon s_n} x^2 \, d\alpha_k(x) = 0} \tag{6}$$

holds, then the central limit theorem holds for the partial sums S_n.

Proof. Like in the proof of the central limit theorem, we use characteristic functions. Let $\phi_k(t)$ denote the characteristic function of the random variable X_k. Then, the characteristic function of the normalized partial sum is given by the product of $\phi_k\left(\frac{t}{s_n}\right)$ and we have to show that

$$\prod_{k=1}^{n} \phi_k\left(\frac{t}{s_n}\right) \to e^{-\frac{t^2}{2}} \qquad \text{for every } t \in \mathbb{R}$$

in the limit $n \to \infty$. For simplicity, we restrict our attention to the case t.

Since the random variables have different distributions, we will need some kind of uniformity; this is precisely what Lindeberg's condition supplies. First, note that it implies

$$\lim_{n\to\infty} \sup_{1 \leq k \leq n} \frac{\sigma_k^2}{s_n^2} = 0.$$

This is because for every $\varepsilon > 0$, we have

$$\frac{\sigma_k^2}{s_n^2} = \frac{1}{s_n^2} \int_{|x|<\varepsilon s_n} x^2 \, d\alpha_k(x) + \frac{1}{s_n^2} \int_{|x|\geq\varepsilon s_n} x^2 \, d\alpha_k(x) \leq \varepsilon^2 + \sum_{l=1}^{n} \frac{1}{s_n^2} \int_{|x|\geq\varepsilon s_n} x^2 \, d\alpha_l(x)$$

and the last term goes to zero by Lindeberg's condition.

Next, we expand the characteristic function using Taylor's theorem: For any function g that is C^2 in a neighborhood of 0, and any number t in that neighborhood, we can find $\xi \in [0, t]$ such that

$$g(t) = g(0) + g'(0)t + g''(0)\frac{t^2}{2} + t(t-\xi)[g''(\xi) - g''(0)].$$

Hence, for the k-th characteristic function, we have

$$\begin{aligned} 1 - \phi_k\left(\frac{t}{s_n}\right) &= -E[iX_k]\frac{t}{s_n} + E[X_k^2]\frac{t^2}{2s_n} - \frac{1}{s_n^2} t(t-\xi) E\Big[X_k^2(e^{i\xi X_k/s_n} - 1)\Big] \\ &= \frac{t^2\sigma_k^2}{2s_n^2} - \frac{1}{s_n^2} t(t-\xi) E\Big[X_k^2(e^{i\xi X_k/s_n} - 1)\Big] \end{aligned} \tag{7}$$

for some $\xi \in [0,t]$ that depends on t and k.

The first consequence of this formula is

$$\lim_{n\to\infty} \sup_{1\le k\le n} \left|1 - \phi_k\left(\frac{t}{s_n}\right)\right| \le \lim_{n\to\infty} 3t^2 \sup_{1\le k\le n} \frac{\sigma_k^2}{s_n^2} = 0.$$

Hence, we can apply the Taylor expansion of the logarithm

$$-\ln(1-x) = x + O(x^2) \implies |\ln x + (1-x)| \le C_q (1-x)^2 \quad \text{for all } |x| < q < 1$$

to write the product of characteristic functions as

$$\prod_{k=1}^n \phi_k\left(\frac{t}{s_n}\right) = \exp\left[\sum_{k=1}^n \ln \phi_k\left(\frac{t}{s_n}\right)\right] = \exp\left[-\sum_{k=1}^n \left(1 - \phi_k\left(\frac{t}{s_n}\right)\right) + R_1(n,t)\right]$$

where the remainder $R_1(n,t)$ satisfies

$$R_1(n,t) \le C_q \sum_{k=1}^n \left|1 - \phi_k\left(\frac{t}{s_n}\right)\right|^2$$

for sufficiently large n. To show that it tends to zero, we estimate

$$C\sum_{k=1}^n \left|1 - \phi_k\left(\frac{t}{s_n}\right)\right|^2 \le C\left(\sup_{1\le k\le n}\left|1 - \phi_k\left(\frac{t}{s_n}\right)\right|\right) \cdot \left(\sum_{k=1}^n \left|1 - \phi_k\left(\frac{t}{s_n}\right)\right|\right).$$

We have already seen that the first factor tends to zero. For the second factor, we have to show that it stays finite. But our goal was to to show that it tends to $t^2/2$ anyway.

In particular, the second consequence of the Taylor expansion of the characteristic function, Eq (7), is that

$$\sum_{k=1}^n \left|1 - \phi_k\left(\frac{t}{s_n}\right)\right| = \sum_{k=1}^n \left(1 - \phi_k\left(\frac{t}{s_n}\right)\right) = \sum_{k=1}^n \frac{t^2\sigma_k^2}{2s_n^2} + R_2(n,t) = \frac{t^2}{2} + R_2(n,t).$$

The remainder $R_2(n,t)$ can be estimated using Lindeberg's condition. For $\varepsilon > 0$ and sufficiently large n, we have

$$\begin{aligned}
|R_2(n,t)| &\le \frac{t^2}{s_n^2}\sum_{k=1}^n E\Big[X_k^2(e^{i\xi(t,k)X_k/s_n} - 1)\Big] \\
&\le \frac{t^2}{s_n^2}\sum_{k=1}^n \int_{|x|<\varepsilon s_n} x^2 \underbrace{\left|e^{i\xi x/s_n} - 1\right|}_{\le e|\varepsilon t| \text{ by Taylor}} d\alpha_k(x) + \frac{t^2}{s_n^2}\sum_{k=1}^n \int_{|x|\ge\varepsilon s_n} x^2 \underbrace{\left|e^{i\xi x/s_n} - 1\right|}_{\le 2} d\alpha_k(x) \\
&\le \frac{t^2}{s_n^2}\sum_{k=1}^n \sigma_k^2 e\varepsilon t + 2\frac{t^2}{s_n^2}\sum_{k=1}^n \int_{|x|\ge\varepsilon s_n} x^2 d\alpha_k(x) \\
&= \varepsilon \cdot et^3 + 2\frac{t^2}{s_n^2}\sum_{k=1}^n \int_{|x|\ge\varepsilon s_n} x^2 d\alpha_k(x).
\end{aligned}$$

By Lindeberg's condition, the second term vanishes in the limit $n \to \infty$. Since $\varepsilon > 0$ was arbitrary, the whole remainder tends to zero.

Taking everything together, we have shown that the product of characteristic functions can be expressed as the exponential of a sum of logarithms, where the latter can be approximated by linear functions whose sum tends to $-t^2/2$, as desired. □

Remark. Ljapunov's condition is the condition that for some $\delta > 0$, we have

$$\boxed{\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \int |x|^{2+\delta} \, d\alpha_k(x) = 0.} \tag{8}$$

It implies implies Lindeberg's condition.

Proof. The proof is rather simple: For any fixed $\varepsilon > 0$, we can move a power of δ outside the integral

$$\int_{|x|\geq \varepsilon s_n} |x|^{2+\delta} \, d\alpha_k(x) \geq \varepsilon^\delta s_n^\delta \int_{|x|\geq \varepsilon s_n} |x|^2 \, d\alpha_k(x).$$

Talking the sum and considering the limit gives Lindeberg's condition

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{|x|\geq \varepsilon s_n} |x|^2 \, d\alpha_k(x) \leq \frac{1}{\varepsilon^\delta} \lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \int |x|^{2+\delta} \, d\alpha_k(x) = 0$$

as desired. □

4 Markov Chains

4.1 Basic notions

In the following, let S be a finite or countably infinite set, the **state space**.

Definition. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(X_0, X_1, X_2, \dots)$ be a sequence of random variables with values in the state space, $X_k : \Omega \to S$. We interpret the index k as "time", so this is a sequence of random variables at a series of discrete time steps. We say that this sequence has the **Markov property** if

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = P(X_{n+1} = x_{n+1} \mid X_n = x_n) \quad (9)$$

whenever the appearing probabilities are well-defined. In other words, the value of the next random variable X_{n+1} is only dependent on the value of the previous random variable X_n, but independent of the earlier history. The conditional probability on the right-hand side is also called the **transition probability**.

In principle, the transition probability could still depend on the time step n, but we will now restrict our attention to sequences where this is not the case:

Definition. A **Markov chain** is a sequence of random variables that has the Markov property and where the transition probability is independent of time. In this case, we introduce the **transition matrix** $\pi(x, y)$ by

$$\boxed{\pi(x, y) := P(X_{n+1} = y \mid X_n = x).}$$

It indicates the probability to "go from x to y". Since the state space is countably, this is indeed a matrix with rows labelled by x and columns labelled by y.

Definition. The **starting distribution** of a Markov chain is the distribution of the initial state, $\nu(x) = P(X_0 = x)$. Usually, one considers Markov chains with the same transition matrix but different starting distributions. We write P_y instead of P when the starting distribution is concentrated at the point y, $\nu(x) = \delta_{x,y}$.

Definition. A square matrix Q with indices in an at most countable set S is called a **stochastic matrix** if

1. All entries are probabilities, $0 \leq Q(x, y) \leq 1$.
2. For every row x, the sum over columns is unity, $\sum_{y \in S} Q(x, y) = 1$.

In other words, the transition probabilities of a Markov chain form a stochastic matrix.

Notation. We also consider the **n-step transition probability**

$$\pi^{(n)}(x, y) = P(X_{k+n} = y \mid X_k = x).$$

For $n = 0$, this is the identity matrix, $\pi^{(0)}(x, y) = \delta_{x,y}$. For $n = 1$, this is the ordinary transition probability, $\pi^{(1)}(x, y) = \pi(x, y)$. In a moment, we will see that this quantitiy is indeed independent of the time index k:

Lemma 4.1 (**Chapman-Kolmogorov equations**). *For any Markov chain, the transition probabilities can be obtained by matrix multiplication*

$$\pi^{(n+m)}(x,y) = \sum_{z\in S} \pi^{(n)}(x,z)\pi^{(m)}(z,y).$$

Proof. We have

$$\begin{aligned} P(X_{n+m} = y \mid X_0 = x) &= \sum_{z\in S} P(X_{n+m} = y, X_m = z \mid X_0 = x) \\ &= \sum_{z\in S} P(X_{n+m} = y \mid X_m = z, X_0 = x) P(X_m = z \mid X_0 = x). \end{aligned}$$

By the Markov property, the first factor in each summand is the n-step transition probability. □

Remark. Conversely, does every stochastic matrix correspond to a Markov chain? Intuitively, this should be true, and under reasonable technical assumptions, the answer will indeed be yes; this is the Ionescu-Tulcea theorem, which we will prove in Section 5.

In consequence, we usually identify a stochastic matrix π with the corresponding Markov chain.

4.2 Examples

Example. A *trivial Markov chain* is a Markov chain where the transition probability does not depend on the past at all, $\pi(x,y) = \pi(x',y) = \tilde{\pi}(y)$ for any states x, x'.

Example. *Random walks.* Let G be any countable group, not necessarily abelian, and μ be a measure on that group. Then, any Markov chain whose state space is that group and whose transition probability has the form

$$\pi(x,y) := \mu(yx^{-1})$$

is called a **random walk**.

For instance, if $G = \mathbb{Z}$, or more generally $G = \mathbb{Z}^d$, and the probability measure is $\mu(e_n) = 1/2n$ for each unit vector $e_n = (0,\dots,0,1,0,\dots 0)$, and $\mu(g) = 0$ otherwise, then this corresponds to walking around on a d-dimensional lattice, where at each turn, we take a step in any of the $2n$ compass directions with equal probability.

Example. *Absorbing random walk.* Consider the state space $S = \{0,1,\dots,N\}$, which corresponds to a linear chain of numbers. An absorbing random walk is a Markov chain where we take one step to the left or to the right with equal probability if both possibilities exist, $\pi(i, i\pm 1) = \frac{1}{2}$ for all states $i \in S = \{1,\dots,N_1\}$, but where we stay in place once one of the two ends of the chain have been reached, $\pi(i,j) = \delta_{ij}$ if $i = 1$ or $i = N$.

Example. *Urn models.* An urn with, say, red and black balls can be represented by a pair $(r, b) \in \mathbb{N}^2$ where r is the number of red and b is the number of black balls in the urn. Taking a ball without putting it back leads to the transition probabilities

$$\pi((r,s),(r-1,s)) = \frac{r}{r+s}, \quad \pi((r,s),(r,s-1)) = \frac{s}{r+s}, \quad \pi((r,s),(r',s') = 0 \text{ otherwise.}$$

Example. *Queue model.* Consider a queue at a counter in a supermarket. At each time step, the queue serves one person, but new persons arrive. We assume that the probability that j persons arrive at the queue in a single time step is p_j, independent of the time. Let X_n denote the number of persons at the queue at time n. Then, the system can be described by a Markov chain with initial distribution and transition matrix

$$\pi(X_1 = i) = p_i, \quad \pi(i,j) = p_{j-(i-1)}.$$

Example. *Time evolution with noise.* A very general model for a physical system is given by the equation

$$X_{n+1} = f(X_n, \xi_n)$$

where X_n is the internal state of the system, ξ_n are i.i.d. random variables that correspond to a random, external influence, and f is a function that determines the internal state at the next time step. This type of model is usually employed when the state space is continuous. For instance, it describes the Brownian motion of a particle.

4.3 Reachability

Definition. Let $x, y \in S$ be two states of a Markov chain. We say that y is **reachable** from x, $x \rightsquigarrow y$, if at least one n-step transition probability is non-zero, $\pi^{(n)}(x,y) \neq 0$.

Definition. We say that two states $x, y \in S$ are in the same **class**, $x \leftrightsquigarrow y$, if y can be reached from x and vice versa.

Remark. The Chapman-Kolmogorov equations show that being in the same class is an equivalence relation. Moreover, for any two states x, y in the same class, we have $\pi^{(m)}(x,y) \neq 0$ for infinitely many natural numbers $m \in \mathbb{N}$.

Definition. A subset of states $A \subset S$ is called **closed** if the only states reachable from this set are in the set itself, that is, for any state $x \in A$ and any state $y \in S \setminus A$, we have $x \not\rightsquigarrow y$.

Definition. A Markov chain is called **irreducible** if all states belong to the same class.

Example. The random walk on $\mathbb{Z}$ or $\mathbb{Z}^d$ is irreducible: Every point can be reached from every other point.

Example. Consider the absorbing random walk on the state space $\{0, 1, \ldots, N\}$ from the previous section. The subsets $J = \{0\}$ and $J = \{N\}$, $J = S$ are closed. All points from the set $S \setminus \{0, N\}$ belong to the same class, but the points 0 and N belong to two additional, different classes. In particular, the absorbing random walk is not an irreducible Markov chain.

4.4 Recurrence and transience

Definition. The **time of first visit** to a state x is denoted by

$$\boxed{T_x := \inf\{n \geq 1 : X_n = x\}.} \tag{10}$$

In particular, the event that a chain starting at x will visit the state y for the first time after exactly n steps has probability

$$P_x(T_y = n) = f^{(n)}(x, y) := \sum_{y_1, y_2, \ldots, y_{n-1} \neq y} \pi(x, y_1)\pi(y_1, y_2) \cdots \pi(y_{n-1}, y). \tag{11}$$

We denote the sum over these probabilities by

$$\boxed{P_x(T_y < \infty) = f(x, y) := \sum_{n=1}^{\infty} f^{(n)}(x, y).} \tag{12}$$

This quantity is smaller than 1, because it is the sum of probabilities of disjoint events. We can interpret it as the probability that the chain visits y after a finite number of steps.

Definition. A state x of Markov chain is called **recurrent** if a chain starting at this state returns almost surely after a finite number of steps, $P_x(T_x < \infty) = 1$. Otherwise, the state is called **transient**.

Lemma 4.2 (**Renewal equation**). *For any two states x, y of a Markov chain, we have*

$$\pi^{(n)}(x, y) = \sum_{k=1}^{n} f^{(k)}(x, y) \cdot \pi^{(n-k)}(y, y) \quad \text{for } n \geq 1. \tag{13}$$

Proof. This corresponds to the disjoint union of events

$$\{X_0 = x, X_n = y\} = \biguplus_{k=1}^{n} \{X_0 = x, X_1 \neq y, X_2 \neq y, \ldots, X_{k-1} \neq y, X_k = y, X_n = y\}.$$

□

Proposition 4.3 (**Recurrence criterion**). *For any Markov chain, we have*

$$\boxed{x \text{ recurrent} \iff \sum_{n=1}^{\infty} \pi^{(n)}(x, x) = +\infty.} \tag{14}$$

The series on the right-hand side can be interpreted as the expected number of number of visits of the state x when starting at x.

Proof. We show that a state is transient if and only if the series on the right-hand side converges.

Since probabilities are always between zero and one, the generating functions

$$\pi(s) := \sum_{n=1}^{\infty} \pi^{(n)}(x, x)s^n \quad \text{and} \quad f(s) := \sum_{n=1}^{\infty} f^{(n)}(x, x)s^n$$

converge absolutely for any parameter $s < 1$ and are monotonically increasing in the parameter s. Moreover, the limit $f = \lim_{s\to 1^-} f(s)$ exists and is equal to the probability of returning to the state in a finite time.

Rewriting the renewal equations (13) in terms of generating functions yields

$$\pi(s) = f(s)(1 + \pi(s)).$$

"$\Longrightarrow$" If the state is transient, $f < 1$, then we rewrite and estimate

$$\pi(s) = \frac{f(s)}{1 - f(s)} \leq \frac{f}{1 - f} < \infty.$$

Letting $s \to 1^-$ shows that the expected number of visits converges absolutely.

"$\Longleftarrow$" If the expected number of visits converges, then its value is equal to $\pi = \lim_{s\to 1} \pi(s)$. Hence, we have

$$f(s) = \frac{\pi(s)}{1 + \pi(s)} \leq \frac{\pi}{1 + \pi} < 1.$$

Letting $s \to 1^-$ shows that $f < 1$ and the state is transient. □

Corollary 4.4. *Recurrence and transience are class properties.*

Proof. If $x \leftrightsquigarrow y$, then there exists natural numbers k, l such that $\pi^{(k)}(x, y) > 0$ and $\pi^{(l)}(y, x) > 0$. By the Chapman-Kolmogorov equations, we have

$$\pi^{(n)}(x, x) \geq \pi^{(k)}(x, y)\pi^{(n-k-l)}(y, y)\pi^{(l)}(y, x) \quad \text{for all } n \in \mathbb{N}, n \geq k + l.$$

Taking the sum over all n shows that x is recurrent if y is recurrent. Similarly for the other direction. □

Proposition 4.5 (**Random walk** on $\mathbb{Z}^d$)**.** *Consider the standard random walk on the lattice $\mathbb{Z}^d$. The n-step transition probability vanishes for odd n and is asymptotically given by*

$$\boxed{\pi^{(n)}(x, x) \simeq \frac{C}{n^{d/2}} \quad \textit{for } n \textit{ even.}} \tag{15}$$

In particular, the random walk is recurrent in dimension $d = 1, 2$, but transient in dimension $d > 3$.

Proof. It is clear that the random walk can only return after an even number of steps.

To calculate the n-step transition probability for even n, we use the Fourier transform. The transition probability only depends on the difference, $\pi(x, y) = p(x - y)$, so the probability $\pi^{(n)}(x, y)$ is given by n-fold convolution of the function $p(z)$ with itself. Convolution of functions corresponds to multiplication of their Fourier transforms, so we have

$$\pi^{(n)}(x, y) = \int_{\mathbb{T}^d} d\xi\, \hat{p}(\xi)^n e^{i2\pi\xi\cdot(x-y)}$$

where

$$\hat{p}(\xi) = \sum_{z\in\mathbb{Z}^d} p(z)e^{-i2\pi\xi\cdot z}$$

is the Fourier transform. We are interested in the case $x = y$, that is we want to estimate the integral

$$\pi^{(n)}(x,x) = \int_{\mathbb{T}^d} d\xi\, \hat{p}(\xi)^n. \tag{16}$$

The Fourier transform of the random walk is given by

$$\hat{p}(\xi) = \frac{1}{2d}\left[e^{-i2\pi\xi_1} + e^{i2\pi\xi_1} + \ldots\right] = \frac{1}{d}[\cos(2\pi\xi_1) + \cos(2\pi\xi_2) + \ldots + \cos(2\pi\xi_d)].$$

In the asymptotic limit $n \to \infty$, the main contributions to the integral (16) are the regions around the points $\xi = (0,\ldots,0)$ and $\xi = (\frac{1}{2},\ldots,\frac{1}{2})$ where the Fourier transform is close to one. After all, if the Fourier transform is smaller than one, $|\hat{p}(\xi)| < 1-\varepsilon$, then its n-th power $(1-\varepsilon)^n$ will decrease exponentially. When n is even, the two regions give equal contributions, so let us focus on the region $|\xi| < \delta$ around the point $\xi = 0$.

In this region, the cosine can bounded by a quadratic functions

$$1 - \cos(2\pi\xi_j) \le 1 - c\xi_j^2 \quad \text{for } |\xi| < \delta.$$

To obtain an accurate estimate for the n-th power, we use the inequality

$$(1-x)^n \le e^{-nx} \text{ for } x \in [0,1],$$

which can be shown e.g. by noting that $1 - x \le e^{-x}$ for $x \in [0,1]$. Then, we have

$$\hat{p}(\xi)^n \le \left(1 - \frac{c}{d}\xi^2\right)^n \le e^{-nC\xi^2}$$

and we can bound the transition probability by a Gaussian integral

$$\pi^{(n)}(x,x) \simeq \int_{|\xi|<\delta} d\xi\, \hat{p}(\xi)^n \le \int_{|\xi|<\delta} d\xi\, e^{-nC\xi^2} \le \frac{C_\delta}{n^{d/2}}.$$

A similar bound can be established in the other other direction. □

4.5 Stopping Times

Definition. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(X_k)_{k=0}^\infty$ be a countable family of random variables. For every natural number n, we define the σ-algebra

$$\mathcal{F}_n := \sigma\text{-algebra generated by the preimages } (X_0, X_1, \ldots, X_n)^{-1}(B) \text{ where } B \subset S^{n+1},$$

which corresponds to those events that can be described by looking only at the first $n+1$ random variables. These σ-algebras from a filtration of the original σ-algebra,

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{A},$$

which is also called the **filtration associated to the random variables** $(X_n)_{n=0}^\infty$.

Definition. Let $(X_n)_{n=1}^\infty$ be a Markov chain. A random variable $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is called a **stopping time** if we have

$$\{T = n\} \in \mathcal{F}_n \quad \text{for all } n \in \mathbb{N}_0.$$

In other words, the question whether $T(\omega) = n$ can be decided by looking only at the values of the first $n+1$ random variables $X_0, X_1, \ldots, X_n$.

Example. The time of first visit to a state, $T_x = \inf\{n \geq 1 : X_n = x\}$, is a stopping time. More generally, the time of first visit to a set of states, $T_A := \inf\{n \geq 1 : X_n \in A\}$ is also a stopping time.

Example. Likewise, the time of second visit to a state, or the first visit to a state x after having visited y are also stopping times.

Example. The time of *last* visit to a state is *not* a stopping time.

Definition. Let T be a stopping time. An event A is called a **pre-T event** if $A \cap \{T = n\} \in \mathcal{F}_n$, that is if its occurrence can be decided by looking only at the random variables before and including the stopping time. The σ-algebra of pre-T events is denoted by $\mathcal{F}_T$.

Example. Consider the time of first visit T_x. The event $A =$ "Visits the state y before returning to the state x for the first time" is a pre-T_x event.

The key feature of stopping times is that the ordinary Markov property (9) not only holds for a fixed time n, but also for any variable stopping time T. In other words, the subsequent variables $X_{T+1}, X_{T+2}, \ldots$ form again a Markov chain with the same transition matrix. This is the statement of the following proposition:

Proposition 4.6 (**Strong Markov property**). *Let T be a stopping time and $A \in \mathcal{F}_T$ a pre-T event. Then, the Markov chain after the time T has lost all memory of its past and starts anew:*

$$\boxed{P(X_{T+n} = x_n, \ldots, X_{T+1} = x_1 \mid X_T = x, A) = P(X_n = x_n, \ldots, X_1 = x_1 \mid X_0 = x),} \qquad (17)$$

assuming that the event that we condition on has non-zero probability, $P(X_T = x, A) > 0$.

Proof. We multiply with the probability $P(X_T = x, A)$ and separate the event according to the different possible values for the stopping time:

$$P(X_{T+n} = x_n, \ldots, X_{T+1} = x_1, X_T = x, A) = \sum_{m=0}^{\infty} P(X_{m+n} = x_n, \ldots, X_{m+1} = x_1, X_m = x, T = m, A).$$

Now, we can use the ordinary Markov property to express the summands on the right-hand side as conditional probabilities and collect the different values of the stopping time again:

$$\begin{aligned} \ldots &= \sum_{m=0}^{\infty} P(X_n = x_n, \ldots, X_1 = x_1 | X_0 = x) P(X_m = x, T = m, A) \\ &= P(X_n = x_n, \ldots, X_1 = x_1 | X_0 = x) P(X_T = x, A). \end{aligned}$$

This concludes the proof. □

The strong Markov property is very useful for gaining a better understanding recurrence. in particular, we obtain the probability distribution for the number of visits:

Proposition 4.7 (**Number of visits**). *Let $V_x = \sum_{n=0}^{\infty} \mathbb{1}_{\{X_n=x\}}$ denote the number of visits to the state x.*

- *If the state is transient, then the number of visits has a geometric distribution*

$$P_x(V_x > k) = p^k \quad \text{with} \quad p = P_x(T_x < \infty).$$

- *If the state is recurrent, then $V_x = \infty$ almost surely.*

Proof. To visit the state more than k times, we have to return to the state once and then visit it more than $k-1$ times. By the strong Markov property, the latter event has the same probability distribution, and we can write

$$P_x(V_x > k) = \sum_{n=1}^{\infty} P_x(T_x = n) \cdot P_x(V_x > k-1) = pP_x(V_x > k-1)$$

where $p = P_x(T_x < \infty)$. Furthermore, we have $P_x(V_x > 0) = 1$. If the state is transient, $p < 1$, it follows that $P(V_x > k) = p^k$. If the state is recurrent, $p = 1$, we conclude $P_x(V_x = k) = 0$ and the number of visits must be infinite almost surely. □

We can use this to prove the fundamental recurrence criterion (14) again:

Corollary 4.8 (**Recurrence criterion**). *For any Markov chain, we have*

$$x \text{ recurrent} \iff E_x[V_x] = +\infty$$

where V_x denotes the number of visits to the state x. In other words, a state is recurrent iff the expected number of visits is infinite.

Proof. The geometric distribution $P_x(V_x > k) = p^k$ with $p < 1$ has a finite expectation value, while the distribution $P_x(V_x = \infty) = 1$ clearly has an infinite expectation value. □

We can show even more: Note only will the chain visit a recurrent state infinitely many times, but it will also visit all reachable states infinitely many times. To prove this, the following Lemma is helpful:

Lemma 4.9 (**Visit before returning**). *Let x be a recurrent state. If there is a positive probability of reaching the state y, $x \rightsquigarrow y$, then there is also a positive probability of reaching the state y before returning to the state x.*

Proof. Let N_y be the event that the chain never visits the state y. Since $x \rightsquigarrow y$, we have $P_x(N_y) < 1$.

Let W_x be the random variable that counts the number of times that the state x is visited before the chain visits y. Since the state x is recurrent, this variable is also well-defined in the case that that the chain never visits y, because then the chain visits the state x infinitely many times almost surely. Put differently, the events N_y and $\{W_x = \infty\}$ differ only by a null set.

Finally, let M_y be the event that the chain does not visit y before returning to the state x. We want to show $P_x(T_x < \infty, M_y) < 1$.

By the strong Markov property, we have the recursive equation

$$P_x(W_x > k) = P_x(T_x < \infty, M_y) P_x(W_x > k-1) \quad \text{for } k \geq 1$$

with starting value $P_x(W_x > 0) = 1$. But we have just argued that the chain cannot return infinitely often without visiting the state y at all, so we have $P_x(W_x = \infty) < 1$. But by the recursive equation, this is only possible if $P_x(T_x < \infty, M_y) < 1$, as desired. □

Corollary 4.10. *If x is recurrent and $x \rightsquigarrow y$, then $y \rightsquigarrow x$. In particular, the two states are in the same class.*

Proof. In the previous lemma, we have shown that the probability of visiting the state y before returning to the state x is positive. But this can only be if the latter is actually reachable from the former, $y \rightsquigarrow x$. □

Proposition 4.11 (**Visits to reachable states**). *If x is recurrent and $x \rightsquigarrow y$, then $P_x(V_y = \infty) = 1$.*

Proof. With the notation from the previous proof, let $p = P_x(T_x < \infty, M_y)$ denote the probability of not visiting the state y before returning to x.

If start at the state x and wait until we have returned to the state x, then we will either have failed to visit the state y, or we will have visited it at least once. By the strong Markov property, we obtain the inequality

$$P_x(V_y > k) \geq p P_x(V_y > k) + (1-p) P_x(V_y > k-1).$$

Here, we have implicitly used that $P_x(V_y > k-1) \leq P_x(V_y > k-1-m)$ for any natural number $m \geq 0$. However, by the previous lemma, the probability $(1-p)$ is positive, so we can divide by it and obtain $P_x(V_y = k) = P_x(V_y = k-1)$. In other words, the number of visits to the other state y is infinite almost surely. □

Corollary 4.12. *If x is recurrent and $x \rightsquigarrow y$, then $P_x(T_y < \infty) = 1$.*

Corollary 4.13. *If x is recurrent and $x \rightsquigarrow y$, then y is recurrent as well. Hence, recurrence is a class property.*

Proof. By the strong Markov property, we can write

$$P_x(V_y = \infty) = P_x(T_y < \infty) P_y(V_y = \infty).$$

But the first two probabilities are equal to one, so the last one has to be equal to one as well. □

Proposition 4.14 (**Recurrence** in a **finite Markov chain**). *In an irreducible Markov chain with finitely many states, all states are recurrent.*

Proof. Let S denote the finite collection of states and let x be any starting state. The total number of visits to any state is infinite almost surely, $P_x(\sum_{y \in S} V_y) = \infty$. Hence, the expectation must be infinite as well, $E_x[\sum_{y \in S} V_y] = \infty$. By the pigeonhole principle, at least one state y has to satisfy $E_x[V_y] = \infty$. Furthermore, we have

$$P_x(V_y = n) = P_x(T_y < \infty) P_y(V_y = n) \quad \text{for } n \geq 1$$

which implies

$$E_x[V_y] = P_x(T_y < \infty) E_y[V_y].$$

Since the Markov chain is irreducible, the probability of reaching the state y from x is positive, and we conclude $E_y[V_y] = \infty$. This means that the state y is recurrent, and by the previous lemmas, this means that the other states are recurrent as well. □

4.6 Invariant measures and recurrence

Definition. Let π be a Markov chain. An **invariant measure** is a measure $\mu : S \to [0, \infty]$ on the set of states that is invariant under transitions,

$$\mu(y) = \sum_{x \in S} \mu(x) \pi(x, y).$$

A **stationary distribution** ν is a stationary measure that is a probability distribution, i.e. that is normalized to one, $\nu(S) = 1$.

In other words, if ν is a stationary distribution and the Markov chain has the starting distribution $P(X_0 = x) = \nu(x)$, then the state in the n-th step will have the same distribution, $P(X_n = x) = \nu(x)$.

Definition. An invariant measure is said to be **trivial** if it is constant $\mu(x) = 0$ or constant $\mu(x) = \infty$. Clearly, a stationary distribution is always nontrivial.

Lemma 4.15. *If an irreducible Markov has a nontrivial invariant measure, then* $0 < \mu(x) < \infty$ *at every state* x*.*

Proof. Since the chain is irreducible, for any two states x, y, there is some step count n such that the former state can be reached from the latter state, $\pi^{(n)}(y, x) > 0$. Then, by the definition of the invariant measure, we have $\mu(x) \geq \mu(y) \pi^{(n)}(y, x)$. If the measure were infinite at some state, $\mu(y) = \infty$, then this inequality shows that it would be infinite at all states. Similarly, if the measure were zero at some state, $\mu(x) = 0$, then it would be zero everywhere. □

It turns out that recurrent Markov chains always have an invariant measure. In fact, we can give an explicit formula for it:

Proposition 4.16 (**Invariant measure** of **recurrent Markov chains**)**.** *Consider an irreducible Markov chain. Let*

$$\boxed{\gamma_x(y) = E_x \left[\sum_{n=1}^{T_x} \mathbb{1}_{\{X_n = y\}} \right]}$$

be the expected number of visits to a state y *before returning to the state* x*. If the Markov chain is* recurrent*, then* γ_x *is the unique invariant measure with* $\gamma_x(x) = 1$*.*

Proof. Invariant measure. To show that γ_x is an invariant measure, we first write the summation in a more uniform way as

$$\sum_{n=1}^{T_x} \mathbb{1}_{\{X_n = y\}} = \sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = y,\, n \leq T_x\}}. \tag{18}$$

The expectation value of this sum is now the sum of probabilities of the individual events. For $n \geq 2$, we have

$$\begin{aligned} P_x(X_n = y, n \leq T_x) &= \sum_{z \neq x} P_x(X_n = y, X_{n-1} = z, n \leq T_x) \\ &= \sum_{z \neq x} P_x(X_n = y, X_{n-1} = z, n-1 \leq T_x) \end{aligned}$$

To test the condition $n - 1 \leq T_x$, we only have to look at the random variables $X_1, \ldots, X_{n-1}$. Hence, we can apply the Markov property to write the right-hand side as

$$P_x(X_n = y, n \leq T_x) = \sum_{z \neq x} \pi(z, y) P_x(X_{n-1} = z, n - 1 \leq T_x).$$

Taking the sum over n, starting at $n = 2$, we obtain

$$\gamma_x(y) - P_x(X_1 = y, 1 \leq T_x) = \sum_{z \neq x} \gamma_x(z)\pi(z, y).$$

Since the Markov chain is recurrent, we have $\gamma_x(x) = 1$. Hence, we can move the second term on the left to the right and see that this is precisely the defining property of an invariant measure.

Uniqueness. Let μ be any invariant measure with $\mu(x) = 1$. We write the equation for invariance in a slightly different way by splitting off the state x:

$$\mu(y) = \sum_{z_1 \neq x} \mu(z_1)\pi(z_1, y) + \mu(x)\pi(x, y) = \sum_{z_1 \neq x} \mu(z_1)\pi(z_1, y) + \pi(x, y).$$

We can iterate this equation (von Neumann series) and use $\mu \geq 0$ to obtain

$$\begin{aligned} \mu(y) &= \sum_{z_1 \neq x} \mu(z_1)\pi(z_1, y) + \pi(x, y) \\ &= \sum_{\substack{z_1 \neq x \\ z_2 \neq x}} \mu(z_2)\pi(z_2, z_1)\pi(z_1, y) + \sum_{z_1 \neq x} \pi(x, z_1)\pi(z_1, y) + \pi(x, y) = \ldots \\ &\geq \sum_{\substack{z_1 \neq x \\ \cdots \\ z_{n-1} \neq x}} \pi(x, z_{n-1}) \cdots \pi(z_2, z_1)\pi(z_1, y) + \ldots + \sum_{z_1 \neq x} \pi(x, z_1)\pi(z_1, y) + \pi(x, y). \end{aligned}$$

However, the right-hand side can be identified with the probabilities of visiting the state y after n steps before returning to x. In other words, we have

$$\mu(y) \geq P_x(X_n = y, n \leq T_x) + \cdots + P_x(X_2 = y, 2 \leq T_x) + P_x(X_1 = y, 1 \leq T_x)$$

Taking the limit $n \to \infty$ and using the previous formula (18) to express this as an expectation value, we see that

$$\mu(y) \geq E_x \left[\sum_{n=1}^{T_x} \mathbb{1}_{\{X_n = y\}} \right] = \gamma_x(y).$$

This means that the measure $\mu - \gamma_x$ is a nonnegative. However, this is clearly an invariant measure with $(\mu - \gamma_x)(x) = 0$. By the previous lemma, the difference must be trivial $\mu - \gamma_x = 0$, as desired. □

To determine those Markov chains that have a stationary distribution, we have to strengthen our notion of recurrence slightly:

Definition. A state is called **positively recurrent** if the chain not only returns almost surely, but also if the expected return time is finite,

$$E_x[T_x] < \infty.$$

It is straightforward to see that a positively recurrent state is also recurrent, because otherwise, the expected return time would be infinite. A state is called **null recurrent** if it is recurrent, but not positively recurrent.

Proposition 4.17 (**Existence** of **stationary distributions**). *For an irreducible Markov chain, the following are equivalent:*

1. *There exists a stationary distribution.*
2. *One state is positively recurrent.*
3. *All states are positively recurrent.*

If these conditions hold, then the stationary distribution is given by

$$\boxed{\nu(x) = \frac{1}{E_x[T_x]}.}$$

Proof. "3 $\implies$ 2" is obvious.

"2 $\implies$ 1" We have to show that the invariant measures γ_x can be normalized. But the sum over all states is simply the expected return time,

$$\sum_{y=x} \gamma_x(y) = \sum_{y=x} E_x\left[\sum_{n=1}^{T_x} \mathbb{1}_{\{X_n=y\}}\right] = E_x\left[\sum_{n=1}^{T_x} \mathbb{1}\right] = E_x[T_x],$$

which is finite by assumption. Hence, the invariant measure can be normalized to a stationary distribution $\nu(y) = \frac{\gamma_x(y)}{E_x[T_x]}$. The indicated formula follows from $\gamma_x(x) = 1$.

"1 $\implies$ 3" Let ν be the stationary distribution. Pick any state x. The measure $\mu(y) = \nu(y)/\nu(x)$ is clearly invariant with $\mu(x) = 1$. Looking at the proof of the previous theorem, we see that this measure bounds the expectation value $\mu(y) \geq E_x\left[\sum_{n=1}^{T_x} \mathbb{1}_{\{X_n=y\}}\right]$. Taking the sum over all states as in the previous case, we obtain

$$\sum_y \mu(y) = \frac{1}{\nu(x)} \sum_y \nu(y) = \frac{1}{\nu(x)} \geq E_x[T_x].$$

Hence, the state x is positively recurrent. □

Remark. In summary, we obtain the following *classification* for irreducible Markov chains:

1. All states are *transient*, $P_x(T_x < \infty) < 1$. There may or may not be an invariant measure, but there is no stationary distribution.

2. All states are *null recurrent*, $P_x(T_x < \infty) = 1$, but $E_x[T_x] = \infty$. There exists a nontrivial invariant measure that is unique up to a prefactor, but it cannot be normalized to a stationary distribution.

3. All states are *positively recurrent*, $E_x[T_x] < \infty$. Then, there exists a unique stationary distribution, which is given by $\nu(x) = \frac{1}{E_x[t_x]}$.

Example. The random walk on $\mathbb{Z}$ does not have a stationary distribution, because the counting measure is the unique nontrivial invariant measure, but cannot be normalized. Hence, this random walk is recurrent, but not positively recurrent.

Example. The random walk on $\mathbb{Z}^3$ is transient, but the counting measure provides a nontrivial invariant measure.

4.7 Convergence to the stationary distribution

We now want to ask what happens when we let a Markov chain run for a very long time. The expectation is that the probability of finding it in a state x will approach the stationary distribution. There is an additional obstruction:

Definition. The **period** of a state x in a Markov chain is defined to be the least common multiple of the step counts for which the chain may return,

$$d_x := \mathrm{lcd}\{n \in \mathbb{N} : \pi^{(n)}(x, x) > 0\}.$$

Definition. A Markov chain is called **aperiodic** if *all* states have period $d_x = 1$.

Example. Consider the Markov chain on $\mathbb{Z}_m$ with transition matrix $\pi(k, l) = \delta_{k+1,l}$. We can represent the states as points on a circle and each transition is simply a single step in counterclockwise direction. In this case, every state has period $d_m = m$. The stationary distribution is simply the uniform distribution on the state. However, it is also clear that the probability $P_k(X_n = l)$ does not converge to the stationary distribution in the limit $n \to \infty$.

Lemma 4.18 (**Period** is a **class property**). *Two states in the same, $x \leftrightsquigarrow y$, always have the same period, $d_x = d_y$.*

Proof. It is enough to show that $d_y \leq d_x$, the other inequality follows by interchanging the two states. Let i be a number of steps such that we can reach y from x, $\pi^{(i)}(x, y) > 0$, and let j be a number of steps so that we can reach x from y, $\pi^{(j)}(y, x) > 0$. Furthermore, let $a = i + j$ denote their sum. Then, we can return to the first state by passing through the second state in this many steps, $\pi^{(a)}(x, x) \geq \pi^{(i)}(x, y)\pi^{(j)}(y, x) > 0$. But this means that the sum a is divisible by the period d_x. Moreoever, for any step number $b \in \mathbb{N}$ where we can return to the first point, we have

$$\pi^{(b)}(x, x) > 0, \text{ hence } \pi^{(a+b)}(y, y) \geq \pi^{(j)}(y, x)\pi^{(b)}(x, x)\pi^{(i)}(x, y) > 0.$$

The lowest common divisor of the sums $a + b$ is still d_x, which means that the period of the second state is a divisor of the period of the first state, as desired. □

Lemma 4.19 (**Aperiodicity criterion**). *A Markov chain is irreducible and aperiodic if and only if, for any two states x, y, there exists a step count $n_0 \in \mathbb{N}$ such that*

$$\pi^{(n)}(x, y) \geq 0 \quad \text{for all } n \geq n_0.$$

Proof. "$\Longleftarrow$" The condition clearly implies that all states are reachable from all other states. Moreover, specializing $y = x$, we have see that the period satisfies $d_x \mid \operatorname{lcd}\{n_0, n_0 + 1\} = 1$ and hence must be equal to one.

"$\Longrightarrow$" Since the Markov chain is irreducible and the transition to another state can be bounded by the transition to the same state, $\pi^{(n)}(x, y) \geq \pi^{(m)}(x, x)\pi^{(n-m)}(x, y)$, it is enough to show the condition for $y = x$.

The argument is essentially one of elementary number theory. Let $n_1, \ldots, n_k$ finitely many step counts that witness the aperiodicity of the state, that is $\pi^{(n_j)}(x, x) > 0$ for each index j and $\operatorname{lcd}\{n_1, \ldots, n_k\} = 1$. Using Euclid's algorithm, we can find potentially *negative* integers $a_1, \ldots, a_k \in \mathbb{Z}$ such that $a_1 n_1 + a_2 n_2 + \ldots + a_k n_k = 1$. Multiplying this equation by every integer in the range $1, 2, \ldots, (n_1 - 1)$ and reducing the resulting equations modulo n_1, we can find *positive* integers $a_{i,j}$ and b_j such that the equations

$$\begin{aligned}
b_1 n_1 + 1 &= a_{2,1} n_2 + \ldots + a_{k,1} n_k \\
b_2 n_1 + 2 &= a_{2,2} n_2 + \ldots + a_{k,2} n_k \\
&\vdots \\
b_{n_1-1} n_1 + (n_1 - 1) &= a_{2,n_1-1} n_2 + \ldots + a_{k,n_1-1} n_k
\end{aligned}$$

hold. Letting $b = \max\{b_1, \ldots, b_{n_1-1}\}$, we see that for any number n of the form $n = bn_1 + m$ with $m \in \mathbb{N}$, we obtain a representation

$$n = bn_1 + m = a_1^{(n)} n_1 + a_2^{(n)} n_2 + \ldots + a_k^{(n)} n_k$$

with *positive* integers $a_j^{(n)}$ by choosing the equation corresponding to the remainder ($m \mod n_1$) and adding a suitable multiple of n_1. For the transition probabilities, this means

$$\pi^{(n)}(x, x) \geq (\pi^{(n_1)}(x, x))^{a_1^{(n)}} \cdots (\pi^{(n_k)}(x, x))^{a_k^{(n)}} > 0$$

Hence, the choice $n_0 = bn_1$ gives the desired result. □

Proposition 4.20 (**Convergence** to the **stationary distribution**). *Consider a Markov chain that is* irreducible, positively recurrent *and* aperiodic. *Then, the probability of reaching any state y after n steps converges towards the stationary distribution,*

$$\lim_{n \to \infty} \pi^{(n)}(x, y) = \nu(y),$$

regardless of the starting state x.

Proof. We prove the statement by means of a *coupling argument*. The idea is to consider two Markov chains $(X_n)_{n=1}^{\infty}$ and $(Y_n)_{n=1}^{\infty}$ with the same transition matrix, but different starting distributions: the first chain starts at a particular initial state x, $P(X_0 = y) = \delta_{x,y}$, whereas

the second chain starts with the stationary distribution $P(Y_0 = y) = \nu(y)$. We will observe both Markov chains in parallel and we will show that they will eventually *meet*, $X_m = Y_m$ for some time step m. At this point, it is clear that the distribution of the state X_n at a later time $n > m$ is the same as the distribution for the state Y_n, which is the stationary distribution.

Coupled chain. To couple the two Markov chains, we simply consider the product of states $((X_n, Y_n))_{n=1}^{\infty}$ on the product probability space $(\Omega \times \Omega', P_X \otimes P_Y)$. This means that at each time step, the variables X_n and Y_n are independent of each other. Hence, this Markov chain has the starting distribution

$$\hat{P}((X_0, Y_0) = (x, y)) = P(X_0 = x)P(Y_0 = y) = \delta_{x,y}\nu(y)$$

and the transition probability

$$\hat{\pi}((x, y), (x', y')) = \pi(x, x')\pi(y, y').$$

Now, the two chains are irreducible and *aperiodic*, hence the previous Lemma 4.19 allows us to conclude that for any two pairs of states $(x, x'), (y, y')$, there is a step count n_0 such that

$$\pi^{(n)}((x, x'), (y, y')) = \pi^{(n)}(x, x')\pi^{(n)}(y, y') > 0 \quad \text{for all } n \geq n_0.$$

Applying the Lemma again, we see that the coupled chain is *irreducible*. (This is where we needed the aperiodicity). Moreover, the distribution $\nu((x, y)) = \nu(x)\nu(y)$ is clearly a stationary distribution for the coupled chain, hence we conclude that the chain is *positively recurrent*.

Meeting time. Let T denote the meeting time, i.e. the time step at which the two chains first meet,

$$T = \inf\{n \in \mathbb{N} : X_n = Y_n\}.$$

We want to show that this time is finite almost surely. To see this, let z be any state and consider the state (z, z) in the coupled chain. Since the coupled chain is *recurrent*, the time of first visit $T_{(z,z)}$ is finite almost surely. But the meeting time is certainly no larger, $T \leq T_{(z,z)}$, hence it is finite almost surely as well.

Switching the chain. We define random variables

$$Z_n = \begin{cases} X_n, & \text{if } n \leq T \\ Y_n, & \text{if } n > T \end{cases}$$

that correspond to switching from the first chain to the second chain when they first meet. We claim that this is a Markov chain with starting distribution $P(Z_0 = y) = \delta_{x,y}$ and transition matrix $\pi(x, y)$. The first claim is obvious. To prove the second claim, we first discriminate on the values of the meeting time:

$$\begin{aligned} P(Z_{n+1} = x_{n+1}, Z_n = x_n, \ldots, Z_0 = x_0) = &\sum_{m=0}^{n} P(Z_{n+1} = x_{n+1}, Z_n = x_n, \ldots, Z_0 = x_0, T = m) \\ &+ P(Z_{n+1} = x_{n+1}, Z_n = x_n, \ldots, Z_0 = x_0, n < T) \end{aligned}$$

Since the event $\{T = m\}$ depends only on the values of X_k and Y_k for time steps $k \leq m$, we can apply the Markov property to the first summands, remembering that $m \leq n$:

$$\begin{aligned} &P(Z_{n+1} = x_{n+1}, Z_n = x_n, \ldots, Z_0 = x_0, T = m) \\ &\quad = P(Y_{n+1} = x_{n+1}, Y_n = x_n, Z_{n-1} = x_{n-1}, \ldots, Z_0 = x_0, T = m) \\ &\quad = \pi(x_n, x_{n+1}) P(Z_n = x_n, \ldots, Z_0 = x_0, T = m). \end{aligned}$$

For the second summand, we use the independence of the chains X_n and Y_n:

$$\begin{aligned} &P(Z_{n+1} = x_{n+1}, Z_n = x_n, \ldots, Z_0 = x_0, n < T) \\ &\quad = P(X_{n+1} = x_{n+1}, X_n = x_n, \ldots, X_0 = x_0, Y_n \neq x_n, Y_{n-1} \neq x_{n-1}, \ldots, Y_0 \neq x_0) \\ &\quad = P(X_{n+1} = x_{n+1}, X_n = x_n, \ldots, X_0 = x_0) \cdot P(Y_n \neq x_n, Y_{n-1} \neq x_{n-1}, \ldots, Y_0 \neq x_0) \\ &\quad = \pi(x_n, x_{n+1}) P(X_n = x_n, \ldots, X_0 = x_0) \cdot P(Y_n \neq x_n, Y_{n-1} \neq x_{n-1}, \ldots, Y_0 \neq x_0) \\ &\quad = \pi(x_n, x_{n+1}) P(X_n = x_n, \ldots, X_0 = x_0, Y_n \neq x_n, Y_{n-1} \neq x_{n-1}, \ldots, Y_0 \neq x_0) \\ &\quad = \pi(x_n, x_{n+1}) P(Z_n = x_n, \ldots, Z_0 = x_0, n < T). \end{aligned}$$

Putting the summands back together proves the claim.

Convergence. Now, we obtain the following marginal distributions for this Markov chain:

$$\begin{aligned} \pi^{(n)}(x, y) = P(Z_n = y) &= P(Z_n = y, n > T) + P(Z_n = y, n \leq T) \\ \nu(y) = P(Y_n = y) &= P(Z_n = y, n > T) + P(Y_n = y, n \leq T). \end{aligned}$$

Subtracting these equations gives

$$|\nu(y) - \pi^{(n)}(x, y)| \leq 2P(n \leq T).$$

Since the chains will meet almost surely, the probability on the right-hand side goes to zero in the limit $n \to \infty$. □

4.8 Reversible Markov chains

Many Markov chain that one encounters have the special property that running them "backward" in time gives the same result as running them "forward" in time. An equivalent definition is the following:

Definition. A Markov chain is called **reversible** if it satisfies the **detailed balance condition**, which means that there is a nontrivial measure $\mu : S \to (0, \infty)$ such that

$$\mu(x)\pi(x, y) = \mu(y)\pi(y, x).$$

Lemma 4.21. *If a Markov satisfies the detailed balance condition, then the measure μ is an invariant measure.*

Remark. For a general Markov chain, an invariant measure does not necessarily satisfy the detailed balance condition.

Proof. A straightforward calculation:

$$\sum_y \mu(y)\pi(y,x) = \sum_y \mu(x)\pi(x,y) = \mu(x)\sum_y \pi(x,y) = \mu(x).$$

□

Lemma 4.22 (**Time reversal**). *Consider a reversible Markov chain started with its stationary distribution. Then, the probability for any sequence of states is the same as the probability for this sequence of states reversed,*

$$P(X_m = x_m, X_{m-1} = x_{m-1}, \ldots, X_1 = x_1, X_0 = x_0) \\ = P(X_m = x_0, X_{m-1} = x_1, \ldots, X_1 = x_{m-1}, X_0 = x_m).$$

Proof. Using the Markov property and the detailed balance condition, we see that the left-hand and right-hand sides are equal to

$$\mu(x_0)\pi(x_0,x_1)\cdots\pi(x_{m-1},x_m) = \pi(x_1,x_0)\pi(x_2,x_1)\cdots\pi(x_m,x_{m-1})\mu(x_m).$$

□

Examples. In various scientific applications, there is the need to sample a probability distribution ν on a very large state space S. The **Monte-Carlo Markov Chain (MCMC)** method is a popular method for doing that.

The idea is that instead of starting with a Markov chain and arriving at a stationary distribution, we start with a stationary distribution v and construct a corresponding Markov chain.

The **Metropolis** algorithm starts with a reference transition probability $\lambda(x,y)$ that is symmetric, i.e. $\lambda(x,y) = \lambda(y,x)$. For instance, this can be a Gaussian distribution, corresponding to a plain random walk in the state space.

The algorithm works as follows: Assume that we are at the state x. To find the next state, we first pick a new state y at random with probability $\lambda(x,y)$. Then, we calculate the ratio $\alpha = \nu(y)/\nu(x)$ of the desired probability distribution. If $\alpha \geq 1$, then the new state is more likely than the old, and we transition to it. Otherwise, we transition to the new state only with probability α, or stay at the old state with probability $1-\alpha$.

This algorithm corresponds to a Markov chain with transition matrix

$$\pi(x,y) = \begin{cases} \lambda(x,y)\min\{1, \frac{\nu(y)}{\nu(x)}\}, & \text{if } x \neq y \\ 1 - \sum_{y\neq x}\pi(x,y), & \text{if } x = y. \end{cases}$$

In fact, this Markov chain is reversible, as it satisfies the detailed balance condition for $x \neq y$:

$$\nu(x)\pi(x,y) = \nu(x)\lambda(x,y)\min\left\{1, \frac{\nu(y)}{\nu(x)}\right\} = \lambda(x,y)\min\{\nu(x),\nu(y)\} = \nu(y)\pi(y,x).$$

Hence, the desired probability distribution ν is the stationary distribution of this Markov chain.

In practical applications, the chain is run up to some time step n. If n is large enough, the chain ends in the state y with probability close to $\nu(y)$. Unfortunately, it is often difficult to give a good estimate for the required running time n. Also, to avoid that the chain is trapped in a small region of the state space, one often runs the chain several times with randomly chosen starting points.

For completeness, here a concrete application in statistical physics. A classical model for magnetism in metals is the *Ising model*. To define the state space, let $I_n = \mathbb{Z}^d \cap [-n, n]^d$ be a lattice of atoms in a very large d-dimensional box. A state is a mapping that assigns a spin $+1$ or -1 to each atom, i.e. the state space is $S = \{+1, -1\}^{I_n}$. The energy of a state $\sigma \in S$ is given by

$$E(\sigma) = -\frac{1}{2} \sum_{\substack{i,j \in I_n \\ |i-j|=1}} J\sigma_i\sigma_j - H \sum_{i \in I_n} \sigma_i.$$

where the first sum is taken over neighboring lattice sites i, j. The real numbers J and H are parameters, corresponding to the spin interaction strength and the external magnetic field. The thermodynamic properties are described by the following probability distribution, the so-called *canonical ensemble*:

$$\nu(\sigma) = \frac{1}{Z} e^{-\beta E(\sigma)} \quad \text{where } Z = \sum_{\sigma \in S} e^{-\beta E(\sigma)}$$

and $\beta = 1/T$ is the inverse of the temperature. From this, we can deduce various measurable quantities. For instance, the total magnetization of the system is the expectation value

$$E_\nu \left[\sum_{i \in I_n} \sigma_i \right]$$

with respect to this probability distribution. If we want to calculate this expectation value numerically, we run into the problem that the state space is very large. This is where the MCMC method can help: We can use a Markov chain to draw samples from the distribution ν and estimate the expectation as an average over the sampled states.

5 Stochastic Processes and Ergodic Theory

5.1 Construction of stochastic processes

In previous sections, we had considered sequences of independent random variables $(X_n)_{n=1}^\infty$, but never clarified whether there actually exists an underlying probability space $(\Omega, \mathcal{A}, P)$ that supports such a countable sequence of independent variables. In this section, we want to solve this outstanding issue and explicitly construct these probability spaces. The first construction will be based on Kolmogorov's consistency theorem. The second construction is known as the Ionescu-Tulcea theorem, it will cover all practically relevant cases.

Definition. Let $(\Omega, \mathcal{A})$ and $(E, \mathcal{E})$ be measurable spaces. An $(E, \mathcal{E})$**-valued stochastic process** is a sequence of random variables $(X_n)_{n=1}^\infty$ such that the maps $X_n : \Omega \to E$ are measurable. Often, we omit mentioning the target space; unless it is clear from the context, we usually refer to the real numbers, $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$, when we speak of a **stochastic process**.

Remark. Any $(E, \mathcal{E})$-valued stochastic process can also be viewed as a map into the set of sequences

$$X : \Omega \to E^{\mathbb{N}}, \quad X(\omega) = (X_n(\omega))_{n=1}^\infty.$$

As explained in Appendix A.1, the set of sequences can be equipped with the product σ-algebra $\mathcal{E}^{\otimes\mathbb{N}}$. It is straightforward to check that the map X is measurable with respect to this σ-algebra if and only if the individual components X_n are measurable. In other words, a stochastic process is identified with a measurable map

$$X : (\Omega, \mathcal{A}) \to (E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}}).$$

Remark. Every stochastic process on the probability space $(\Omega, \mathcal{A}, P)$ gives rise to a probability distribution on sequences, $(E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}}, P \circ X^{-1})$. Conversely, every probability measure on the space of sequences gives rise to a stochastic process, by choosing $\Omega = E^{\mathbb{N}}$ and letting X be the identity map. Hence, to prove the existence of i.i.d. random variables, we have to construct product measures on the space of sequences.

Remark. Given a probability distribution μ on the sequence space $(E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}})$, we can define the **finite-dimensional distributions**

$$\mu_n := \mu \circ \rho_n^{-1}$$

where $\rho_n : E^{\mathbb{N}} \to E^n$, $\rho_n((x_k)_{k=1}^\infty) = (x_1, \dots, x_n)$ is the projection onto the first n coordinates. These distributions are consistent in the following sense:

Definition. Let $(E, \mathcal{E})$ be a measurable space. A sequence $(\mu_n)_{n=1}^\infty$ of probability measures on the spaces $(E^n, \mathcal{E}^{\otimes n})$ is called **consistent** if

$$\boxed{\mu_n = \mu_{n+1} \circ \phi_n^{-1} \quad \text{for all } n \in \mathbb{N}}$$

where $\phi_n : E^{n+1} \to E^n$, $\phi_n(x_1, \dots, x_n, x_{n+1}) = (x_1, \dots, x_n)$ is the projection that discards the last coordinate.

Now, the question is whether any consistent family can also be obtained as the family of finite-dimensional distributions on the sequence space. In general, this is not the case. For instance, we need to require that the space E is a nice topological space:

Theorem 5.1 (**Kolmogorov consistency theorem**). *Consider the euclidean space* $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}^{\otimes d})$. *Then, given any consistent family* $(\mu_n)_{n=1}^{\infty}$ *of probability distributions, there exists a unique measure* μ *on the sequence space* $(E^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$ *such that the distributions are the finite-dimensional distributions of this measure,* $\mu_n = \mu \circ \rho_n^{-1}$.

Proof. We want to apply the Hahn-Kolmogorov extension theorem A.4.

In the space of sequences, let $\mathcal{F}_n$ be the σ-algebra generated by the projection on the first n-coordinates, $\mathcal{F}_n = \rho_n^{-1}(\mathcal{E}^{\otimes n}) \subset \mathcal{E}^{\otimes \mathbb{N}}$. Then, the countable union $\mathcal{A} = \bigcup_{n \in \mathbb{N}} \mathcal{F}_n$ is an algebra that generates the σ-algebra of the sequence space $\mathcal{E}^{\otimes \mathbb{N}}$.

Since the projection ρ_n^{-1} is a bijective mapping between the σ-algebras $\mathcal{F}_n$ and $\mathcal{E}^{\otimes n}$, there is a unique measure $\tilde{\mu}_n$ on $\mathcal{F}_n$ with the property that $\mu_n = \tilde{\mu}_n \circ \rho_n^{-1}$. Thanks to the assumption of consistency, these measures are all compatible with each other, that is any two measures $\tilde{\mu}_n$ and $\tilde{\mu}_m$ agree on the σ-algebra $\mathcal{F}_{\min\{n,m\}}$. Together, they define a content μ on the algebra $\mathcal{A}$. All we have to show is that this content is σ-additive; then, the Hahn-Kolmogorov extension theorem will give us a unique measure on the whole σ-algebra that extends this content.

We can prove σ-additivity by showing continuity: If $B_1 \supset B_2 \supset \dots$ is a monotonically decreasing sequence of sets in $\mathcal{A}$ with $\mu(B_n) \not\to 0$, then we need to show that the intersection $B = \bigcap_{n=1}^{\infty} B_n$ is non-empty. Now, if all sets in this sequence are contained in one of the σ-algebras $\mathcal{F}_N$, then we are done, because the restriction of μ onto this σ-algebra is a measure. If this is not the case, then by repeating some elements in the sequence, we may assume that $B_n \in \mathcal{F}_n$. Consequently, there exists sets $A_n \in \mathcal{E}^{\otimes n}$ such that $B_n = \rho_n^{-1}(A_n)$. Monotonicity of the sequence is equivalent to the statement that $A_n \times E \supset A_{n+1}$.

To prove that the intersection is non-empty, we want to find a sequence $x = (x_1, x_2, \dots) \in E^{\mathbb{N}}$ such that each finite prefix is contained in the corresponding set, $(x_1, \dots, x_n) \in A_n$. Then, it follows that $x \in \bigcap_{n=1}^{\infty} B_n$, and the intersection cannot be empty.

To construct the desired sequence, we use that every finite measure on a euclidean space is *inner regular*, which means that

$$\mu_n(B) = \sup\{\mu_n(K) : K \subset B, K \text{ compact}\} \quad \text{for all Borel sets } B \subset \mathbb{R}^{dn}.$$

See also Section A.2. In other words, for each index n, we can find a compact subset $K_n \subset A_n$ such that

$$\mu_n(K_n) \geq \mu_n(A_n) - 2^{-n}\delta/2.$$

However, the preimages $\rho_n^{-1}(K_n)$ are not necessarily contained in each other, so let us consider the intersections

$$L_n := \bigcap_{m=1}^{n} K_m \times E^{n-m} \subset K_n \subset A_n.$$

These sets are also compact and satisfy $L_n \times E \supset L_{n+1}$. Additionally, we have

$$\begin{aligned}\mu_n(A_n) &= \mu_n\left(\bigcap_{m=1}^{n} A_m \times E^{n-m}\right) = \mu_n\left(\bigcap_{m=1}^{n} (K_m \uplus (A_m \setminus K_m)) \times E^{n-m}\right) \\ &\leq \sum_{m=1}^{n} \mu_m(A_m \setminus K_m) + \mu_n\left(\bigcap_{m=1}^{n} K_m \times E^{n-m}\right) \leq \delta/2 + \mu_n(L_n).\end{aligned}$$

In other words, the measures of the sets L_n also converge to a non-zero value, $\mu_n(L_n) \geq \delta/2 > 0$.

To find the desired sequence, we can now apply compactness and use a diagonal argument. Let $(x_n)_{n=1}^{\infty}$ be a sequence of points in the sequence space $x_n \in E^{\mathbb{N}}$ such that $x_n \in \rho_n^{-1}(L_n)$. Since the set L_1 is compact, we can choose a subsequence $(x_{m_1(n)})_{n=1}^{\infty}$ such that the first coordinate $\rho_1(x_{m_1(n)})$ converges in L_1. Since the set L_2 is compact, we can choose a subsequence of this one so that the second coordinate converges as well. Repeat this process. Finally, consider the diagonal sequence $y_n = x_{m_n(n)}$. It converges at all coordinates, so there is a limit $y = \lim_{n\to\infty} y_n$ with the property that $\rho_n(y) \in L_n$ for all indices $n \in \mathbb{N}$. Hence, it lies in the intersection $\bigcap_{n=1}^{\infty} A_n$, which means that the latter must be non-empty, as desired. □

Corollary (**Existence of i.i.d. random variables**). Let P be any probability measure on the real numbers $(\mathbb{R}, \mathcal{B})$. Then, the product measures $Q_n := P^{\otimes n}$ are a consistent family of probability measures on the spaces $(\mathbb{R}^n, \mathcal{B}^{\otimes n})$. Since the space of real numbers is a polish space, Kolmogorov's consistency theorem applies, and we obtain a sequence of random variables $(X_n)_{n=1}^{\infty}$ that are independent and indvidiually distributed according to the distribution P.

If we want to drop the assumption that the space E is a euclidean space $\mathbb{R}^d$, then we will have to be more restrictive about the measures allowed. This is the content of the Ionescu-Tulcea theorem, which we will prove in the following paragraphs.

Definition. Let $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ be two measurable spaces. A **Markov kernel** from E_1 to E_2 is a map

$$K : E_1 \times \mathcal{E}_2 \to [0, 1]$$

such that for each point $x \in E_1$, the map $K(x, \cdot)$ is a probability measure, and for each set $A \in \mathcal{E}_2$, the map $K(\cdot, A)$ is measurable.

If we think of the measurable spaces as state spaces, then a Markov kernel assigns to each state x the probability $K(x, A)$ that the next state will be in the set A. If we now pick the first state randomly and the next state according to the Markov kernel, then we will obtain a probability for pairs of states (x, y). The following definition captures this intuition:

Definition. Given a Markov kernel K from E_1 to E_2 and a probability measure μ on the space E_1, we can define a measure on the product space $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ by

$$\boxed{(\mu \otimes K)(A) := \int_{E_1} K(x, A_x)\, d\mu(x).} \tag{19}$$

Here, $A_x = \{y \in E_2 : (x, y) \in A\}$ is the section of the set A at the first coordinate x.

Proof. It is not immediately obvious that this is well-defined.

First, note that the sections A_x are measurable, because this is true for direct products $A = A_1 \times A_2$ and extends to the generated σ-algebra.

Second, the map $x \mapsto K(x, A_x)$ is always measurable. This is again true for direct products $A = A_1 \times A_2$ and extends to the generated σ-algebra. Since the measure μ is a probability measure and the Markov kernel is bounded, it follows that this map is integrable.

Lastly, σ-additivity follows from the σ-additivity of the Markov kernel and Lebesgue's dominated convergence theorem. □

Remark. A variant of Fubini's theorem also holds for product measures constructed in this manner. We will not prove this here.

The following theorem now gives us the existence of stochastic processes for all practically relevant problems:

Proposition 5.2 (**Ionescu-Tulcea**). *Let $(E, \mathcal{E})$ be a measurable space. Let ν be a probability measure on this space and for each $n \in \mathbb{N}$, let K_n be a Markov kernel from $(E^n, \mathcal{E}^{\otimes n})$ to $(E, \mathcal{E})$. Inductively define measures μ_n on the space $(E^n, \mathcal{E}^{\otimes n})$ by*

$$\mu_1 = \nu, \quad \mu_{n+1} = \mu_n \otimes K_n.$$

Then, there exists a unique measure μ on the sequence space $(E^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$ such that these measures are induced by the projections onto the first coordinates, $\mu_n = \mu \circ \rho_n^{-1}$.

Remark. If a sequence starts with the points $(x_1, x_2, \ldots, x_n)$, then the Markov kernel $K_n((x_1, \ldots, x_n), \cdot)$ gives the probability measure for the next point x_{n+1}. The purpose of Ionescu-Tulcea's theorem is to extend this to a probability measure for infinite sequences.

Remark. ▷ todo: Markov kernel $K_{n,m}$ for going from n to m. ◁

Proof. The beginning of the proof is similar to the beginning of the proof of Kolmogorov's consistency theorem 5.1: First, we find that there is a unique content μ on the sequence space such that $\mu(\rho_n^{-1}(A)) = \mu_n(A)$ for each set in the n-fold product, $A \in \mathcal{E}^{\otimes n}$. Then, we want to show continuity of this content. For this, we have to show that if $A_n \in \mathcal{E}^{\otimes n}$ is a sequence of sets with $A_n \times E \supset A_{n+1}$ whose measures do not converge to zero, $\mu_n(A_n) \geq \delta > 0$, then there exists an element $x \in \bigcap_{n=1}^{\infty} \rho_n^{-1}(A_n)$, so that the intersection is not empty.

In the case of Kolmogorov's consistency, we have used compactness to show the existence of the desired sequence. Here, we have to proceed differently.

Consider the following functions

$$f_{n,m} : E^n \to [0,1], \quad f_{n,m}(y) := \int \mathbb{1}_{A_m}(y,z)\, K_{n,m}(y, dz) \quad \text{for } n, m \in \mathbb{N}, m \geq n+1.$$

We set $f_{n,n}(y) := \mathbb{1}_{A_n}(y)$. These functions represent the conditional probability that a point $x \in E^m$ falls in the set A_m under the condition that it starts with the prefix $y \in E^n$, i.e. $x = (y, z)$ with suffix $z \in E^{m-n}$. Since the sets were contained in each other, $A_m \times E \supset A_{m+1}$, this sequence of conditional probabilities is monotonically decreasing

$$1 \geq f_{n,n}(y) \geq f_{n,n+1}(y) \geq \ldots \geq 0 \quad \text{for all } n \in \mathbb{N}, y \in E^n.$$

Moreover, choosing the prefix at random, we obtain the total probability,

$$\int f_{n,m}(y)\, d\mu_n(y) = \mu_m(A_m).$$

We now proceed to construct the sequence $(x_1, x_2, \dots)$ inductively. First, consider the total probability above specialized to $n = 1$. By assumption, the right-hand side converges to a non-zero value in the limit $m \to \infty$. By the Lebesgue dominated convergence theorem, the sequence of functions $f_{1,m}$ cannot converge to 0 almost everywhere, so there must exist a value $x_1 \in E$ such that $\lim_{m\to\infty} f_{1,m}(x_1) > 0$. By monotonicity, we have $\mathbb{1}_{A_1}(x_1) = f_{1,1}(x_1) > 0$, so the prefix is contained in the set, $x_1 \in A_1$.

For the induction step, we use Fubini's theorem, which implies that the functions satisfy the recursive equation

$$f_{n,m}(y) = \int f_{n+1,m}(y, x_{n+1})\, K_n(y, dx_{n+1}).$$

Given a sequence $y = (x_1, x_2, \dots, x_n)$ with $\lim_{m\to\infty} f_{n,m}(y) > 0$, the left-hand side converges to a non-zero value in the limit $m \to \infty$. By Lebesgue's dominated convergence theorem and monotonicity, there must exist a value $x_{n+1} \in E$ such that $\lim_{m\to\infty} f_{n+1,m}(y, x_{n+1}) > 0$ as well. As before, this shows $(x_1, \dots, x_n, x_{n+1}) \in A_{n+1}$, as desired. □

Example. ▷ todo ◁ i.i.d. variables.

Example. Markov chains.

5.2 Stationary Processes and Ergodicity

Definition. A stochastic process $(X_n)_{n=1}^{\infty}$ is called a **stationary process** if the shifted process $(X_{n+1})_{n=1}^{\infty}$ has the same probability distribution.

Example. Every i.i.d. process is a stationary process. In particular, a Bernoulli process is stationary.

Remember that a stochastic process can also be viewed as a measurable map $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ to the space of sequences. If we let

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

denote the shift operator on the space of sequences, then a process X is stationary if and only if the shift θ leaves the probability distribution $P \circ X^{-1}$ invariant, that is $P \circ X^{-1} = P \circ X^{-1} \circ \theta^{-1}$. This motivates the following definition:

Definition. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. A measurable map $T : \Omega \to \Omega$ is said to be **measure-preserving** if

$$\mu(T^{-1}(A)) = \mu(A) \quad \text{for all } A \in \mathcal{A}.$$

A **measure-preserving system** is a quadruple $(\Omega, \mathcal{A}, \mu, T)$ where $(\Omega, \mathcal{A}, \mu)$ is a probability space, $\mu(\Omega) = 1$, and T is a measure-preserving transformation.

In other words, every stationary process $(X_n)_{n=1}^{\infty}$ corresponds to a unique measure-preserving system on the sequence space, $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, P \circ X, \theta)$.

Remark. Conversely, given a measure-preserving system $(\Omega, \mathcal{A}, \mu, T)$ and a random variable $Y : \Omega \to \mathbb{R}$, we can define a stochastic process by $X_n(\omega) := Y(T^n\omega)$. This process is *stationary*, because

$$\begin{aligned}\mu((X_2, X_1, \ldots, X_{n+1}) \in B) &= (\mu \circ T^{-1})((X_1, X_2, \ldots, X_n) \in B) \\ &= \mu((X_1, X_2, \ldots, X_n) \in B),\end{aligned}$$

so the finite-dimensional and hence also the infinite distributions agree. If the system is the sequence space associated to a stochastic process, $(\Omega, \mathcal{A}, \mu, T) = (\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, P \circ X^{-1}, \theta)$, then the original process is recovered by choosing $Y = \pi_1$ the projection onto the first coordinate.

In other words, the concepts "stationary process" and "measure-preserving system" are interchangeable. We can focus on concrete sequence spaces that correspond to stochastic processes,

$$(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}, P \circ X^{-1}, \theta, \pi_1),$$

or we can consider abstract measure-preserving systems with an additional variable $Y : \Omega \to \mathbb{R}$,

$$(\Omega, \mathcal{A}, \mu, T, Y).$$

The former is an instance of the latter, and the latter can always be mapped to the former. in the following, we will adopt mainly the abstract viewpoint.

We now consider some examples of measure-preserving systems:

Example. *Circle rotations.* Consider the unit circle in the complex plane, $\Omega = S^1 \subset \mathbb{C}$, with the (induced) Lebesgue measure. For each real number $\alpha \in [0, 1]$, the rotation

$$T_\alpha : S^1 \to S^1, \quad T_\alpha(z) = e^{2\pi i\alpha} z$$

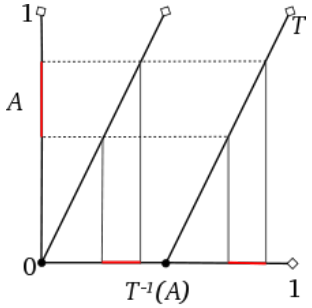
is measure-preserving. An equivalent way of writing this map is to view the circle as the quotient $S^1 \cong \mathbb{R}/\mathbb{Z}$. Then, the rotation corresponds to a linear shift

$$\tilde{T}_\alpha : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}, \quad \tilde{T}_\alpha(x) = x + \alpha \mod \mathbb{Z}.$$

Example. Consider the unit interval $\Omega = [0, 1]$. The transformation

$$T(x) = (2x \mod \mathbb{Z}) = \begin{cases} 2x, & \text{if } 0 \le x < \frac{1}{2} \\ 2x - 1 & \text{if } \frac{1}{2} \le x \le 1 \end{cases}$$

is neither continuous nor bijective. However, it is measure-preserving. This may seem odd, because if I is a small interval, then its image $T(I)$ has twice the length, $|T(I)| = 2|I|$. If we consider the preimage $T^{-1}(J)$ of a small interval J, then the preimage consists of *two* parts of half the length, so the map is measure-preserving, indeed. The following picture (Wikipedia) illustrates this:



Example. *Stationary Markov chains.* Any positively recurrent Markov chain that is started in its stationary distribution ν, so that $P(X_1 = x) = \nu(x)$, is a stationary process. After all, the probability for the second state is the same, $P(X_2 = x) = \nu(x)$, and the finite-dimensional probabilities are equal

$$\begin{aligned} P(X_2 = x_1, X_3 = x_2, \ldots, X_{n+1} = x_n) &= P(X_2 = x_1) \prod_{k=1}^{n} \pi(x_k, x_{k+1}) = \nu(x_1) \prod_{k=1}^{n} \pi(x_k, x_{k+1}) \\ &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n). \end{aligned}$$

Definition. A set A is called **T-invariant** if the transformation maps it to itself, that is $T^{-1}A = A$.

Definition. A measure-preserving system $(\Omega, \mathcal{A}, P, T)$ is called **ergodic** if the following implication holds

$$\boxed{A \text{ is } T\text{-invariant } \implies P(A) = 0 \text{ or } P(A) = 1.}$$

In other words, this means that every T-invariant set either contains almost no points, or almost all points.

A very nice property that implies ergodicity is the following:

Definition. A measure-preserving system $(\Omega, \mathcal{A}, P, T)$ is called **mixing** if

$$\boxed{\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B)} \tag{20}$$

for all measurable sets A, B. We can interpret A and $T^{-n}B$ as two events that separated by a time difference n. Mixing means that these events will become independent in the asymptotic limit of large time differences.

Lemma 5.3. *If a measure-preserving system is mixing, then it is also ergodic.*

Proof. Let A be a T-invariant set. Then, we have $T^{-n}A = A$, so

$$P(A) = \lim_{n\to\infty} P(A \cap A) = \lim_{n\to\infty} P(A \cap T^{-n}A) = P(A)P(A).$$

It follows that $P(A) = 0$ or 1. □

Example. The measure-preserving system associated to a sequence of i.i.d. random variables $(X_n)_{n=1}^{\infty}$ is mixing. It is sufficient to check this for events A and B that only depend on the

variables $X_1, \ldots, X_n$. In this case, the events A and $T^{-m}B$ are independent for any $m > n$, because the latter event only depends on the values of the variables $X_m, \ldots X_{m+n}$.

Example. ▷ todo: irreducible, stationary Markov chains are mixing ◁

Example. Consider the unit interval $\Omega = [0, 1]$ and the map

$$T(x) = 2x \mod \mathbb{Z}.$$

This measure-preserving system is actually equivalent to a Bernoulli system with probability $1/2$, and hence mixing.

Proof. To see this, let $x \in [0, 1]$ be a number and consider the expansion in binary digits $x = 0.b_1b_2b_3\ldots$. This expansion is unique except when the number x is an integer multiple of a negative power of two, $x = k2^{-n}$. But these cases have measure zero, so we can exclude them, and obtain a map into a sequence space, $([0, 1], \mathcal{B}) \to (\{0, 1\}^{\mathbb{N}}, \{0, 1\}^{\otimes\mathbb{N}})$. Moreover, the Lebesgue measure is mapped to a measure where every digit has equal probability $1/2$. Finally, the operation T maps the number x to a number with binary digit expansion $T(x) = 0.b_2b_3\ldots$. In other words, it corresponds to the shift operator, and the system is equivalent to a Bernoulli system. □

Mixing is a very strong property. Not every ergodic system is mixing. The simplest example is given by circle rotations.

Proposition 5.4 (Circle rotations, ergodic, not mixing). *Consider the unit circle $\Omega = S^1$ with the Lebesgue measure. For each real number $\alpha \in [0, 1]$, consider the rotation*

$$T_\alpha : S^1 \to S^1, \quad T_\alpha(z) = e^{2\pi i\alpha}z.$$

Then, we have

$$\boxed{T_\alpha \text{ ergodic} \quad \Longleftrightarrow \quad \alpha \text{ is irrational.}}$$

Moreover in the case where T_α is ergodic, it is not mixing.

Proof. As a measure space, we can identify the circle with the unit interval, $S^1 \cong \mathbb{R}/\mathbb{Z} \cong [0, 1]$, and the operation with the shift $T_\alpha(x) = x + \alpha \mod \mathbb{Z}$.

If α is rational, then it is not difficult to construct an invariant set with non-trivial measure. If $\alpha = p/q$ with $p, q \in \mathbb{N}$, then the shift is periodic, $T^q = 1$. If $B = [0, \varepsilon]$ is a small interval, say $\varepsilon = 1/(2q)$, then the sets $T^{-k}B$ are disjoint for $k = 0 \ldots q - 1$ and the set

$$A = B \cup T^{-1}B \cup T^{-2}B \cup \ldots \cup T^{-(q-1)}B$$

is T-invariant. It has measure $\mu(A) = q\mu(B) = 1/2$. Hence, the system is not ergodic.

If α is irrational, we first establish that the orbit of any point $z \in S^1$ is dense. Without loss of generality, we show this for the point $z = 1$. If we set $\xi = e^{2\pi i\alpha}$, then the orbit of this point is given by $\{1, \xi, \xi^2, \ldots\}$. Let $\varepsilon > 0$. Divide the circle into regions of length ε. By the pigeonhole principle, there must exist some integers $n > m > 0$ such that two orbit points fall into the same region, that is

$$|\xi^n - \xi^m| < \varepsilon,$$

which implies

$$|\xi^{n-m} - 1| < \varepsilon.$$

But since the angle was irrational, we have $\xi^{n-m} \neq 1$. Hence, any other point $z' \in S^1$ on the circle can be approximated within ε by an integer power of this number

$$|\xi^{k(n-m)} - z'| < \varepsilon.$$

This shows that the orbit is dense.

Now, let A be a measurable, T-invariant set. Assume that $0 < \mu(A) < 1$. By measure theory, we can approximate it by a finite disjoint union of intervals I_k, so that $\mu(A \triangle \uplus_{k=1}^{N} I_k) < \varepsilon\mu(A)$ with $\varepsilon > 0$ chosen later. Without loss of generality, we may assume that all the intervals have size $\mu(I_k) < \varepsilon$. By the pigeonhole principle, the intersection of the set A with one of the intervals, say I_1, must be almost the whole of the interval,

$$\mu(A \cap I_1) \geq (1 - 2\varepsilon)\mu(I_1),$$

because otherwise, we would have

$$(1-\varepsilon)\mu(A) \leq \mu\left(A \cap \biguplus_{k=1}^{N} I_k\right) = \sum_{k=1}^{N} \mu(A \cap I_k) < (1-2\varepsilon)\sum_{k=1}^{N} \mu(I_k) \leq (1-2\varepsilon)(1+\varepsilon)\mu(A).$$

Choosing ε small enough, this would be a contradiction to $\mu(A) > 0$.

Now, since the orbit of the action T is dense, we can use disjoint translates of the interval I_1 to cover the unit circle. In particular, we can find integers $n_1, \ldots, n_K$ such that the translates $T^{-n_k} I_1$ are disjoint and

$$\mu(I_1 \cup T^{-n_1} I_1 \cup \ldots \cup T^{-n_K} I_1)) \geq 1 - \varepsilon.$$

However, the set A is T-invariant, so the intersection with this covering has size at least

$$\begin{aligned}
\mu(A) &\geq \mu(A \cap (I_1 \cup T^{-n_1} I_1 \cup \ldots \cup T^{-n_k} I_1)) \\
&= \sum_{k=1}^{K} \mu(A \cap T^{-n_k} I_1) = \sum_{k=1}^{K} \mu(T^{-n_k} A \cap T^{-n_k} I_1) = \sum_{k=1}^{K} \mu(A \cap I_1) \\
&\geq (1-2\varepsilon)\sum_{k=1}^{K} \mu(I_1) \geq (1-2\varepsilon)(1-\varepsilon).
\end{aligned}$$

Letting $\varepsilon \to 0$, we conclude that is $\mu(A) = 1$, so the system is ergodic.

To show that the system is not mixing, consider the interval $A = B = [0, 1/2]$. Since the orbit of any point is dense, there exist infinitely many indices $n_1, n_2, \ldots$ such that $|T^{n_k}(0)| < \varepsilon$. In other words, the interval $T^{-n_k} B$ almost completely overlaps with the original interval B, which means

$$P(A \cap T^{-n_k} B) = 1/2 - \varepsilon > 1/4 = P(A)P(B).$$

This proves that the system cannot be mixing. □

5.3 The Ergodic Theorem

Theorem 5.5 (**Birkhoff's ergodic theorem**). *Let $(\Omega, \mathcal{A}, P, T)$ be a measure-preserving system and $X \in L^1$. Then, the following averages converge almost surely*

$$\lim_{n\to\infty} \frac{X(\omega) + X(T(\omega)) + \ldots + X(T^n(\omega))}{n} = Y(\omega)$$

to a random variable $Y \in L^1$ with $E[X] = E[Y]$.

Corollary 5.6 (**Strong Law of Large Numbers**). *Let $(X_n)_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with mean $E[X_k] = \mu$. Then, we have*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \quad \textit{almost surely.}$$

Proof. The i.i.d. sequence of random variables corresponds to a measure-preserving system that is mixing, hence ergodic. By Birkhoff's ergodic theorem, the averages converge almost surely to a random variable $Y(\omega)$. Since the system is ergodic, this variable is constant almost everywhere. Since $\mu = E[X_k] = E[Y]$, we conclude that it is almost everywhere equal to the mean μ. □

Proof of the ergodic theorem. Without loss of generality, we may assume that the random variable X is positive, $X \geq 0$, otherwise we express it as a difference of a positive and a negative part $X = X_+ - X_-$.

Consider the random variables

$$\overline{X}(\omega) = \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(T^k\omega), \quad \underline{X}(\omega) = \liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(T^k\omega).$$

These variables are invariant under T, that is $\overline{X} \circ T = \overline{X}$, and likewise for the limes inferior. To prove the theorem, it is enough to show the inequalities

$$\int \underline{X}\, dP \geq \int X\, dP \geq \int \overline{X}\, dP.$$

This establishes both that $\overline{X}$ is integrable and that $\underline{X} = \overline{X}$ almost everywhere, i.e that the averages converge almost everywhere. We will only show the second inequality, the first inequality is analoguous.

For $C > 0$, consider the function $\overline{X}^C = \min\{\overline{X}, C\}$ that has been cut off from above. Let $\varepsilon > 0$. The key idea to the proof is to consider, for each point $\omega \in \Omega$, the smallest number of steps needed for the average to come close to the limes superior (or the cut-off),

$$n(\omega) := \min\left\{ n : \frac{1}{n} \sum_{k=0}^{n-1} X(T^k\omega) \geq \overline{X(\omega)}^C - \varepsilon \right\}.$$

This function is measurable. If it were bounded, $n(\omega) \leq N_\varepsilon$, then we could argue that the sequence of values $X(\omega), X(T\omega), \ldots, X(T^m\omega)$ can be subdivided into sections of length at

most N_ε,

$$\underbrace{X(\omega), X(T\omega), X(T^2\omega), \dots, X(T^{n(\omega)-1}\omega)}_{n(\omega)\le N_\varepsilon \text{ values}}, \underbrace{X(T^{n(\omega)}\omega), X(T^{n(\omega)+1}\omega), X(T^{n(\omega)+2}\omega), \dots}_{n(T^{n(\omega)}\omega)\le N_\varepsilon \text{ values}}, \dots, X(T^m\omega),$$

and the average over each section comes close the limes superior. Then, the total average over m values is a weighted average of these sections, plus a remainder, which contribues at most N_ε values. The latter is irrelevant in the limit $m \to \infty$, and the former stays close to the limes superior.

While the step size $n(\omega)$ does not need to be bounded in general, we can certainly find a number N_ε and a set A that is small, $P(A) < \varepsilon$, such that the function is bounded outside this set, $n(\omega) \le N_\varepsilon$ for $\omega \in A^c$. It turns out that this is enough for the argument to work. To see this, consider the adjusted functions

$$\tilde{X}(\omega) = \begin{cases} X(\omega), & \text{if } \omega \in A \\ C, & \text{if } \omega \in A^c, \end{cases} \quad \text{and} \quad \tilde{n}(\omega) = \begin{cases} n(\omega), & \text{if } \omega \in A \\ 1, & \text{if } \omega \in A^c. \end{cases}$$

Clearly, the function $\tilde{n}(\omega)$ is now bounded. Also, it satisfies

$$\frac{1}{\tilde{n}(\omega)} \sum_{k=0}^{\tilde{n}(\omega)-1} \tilde{X}(T^k\omega) \ge \overline{X(\omega)}^C - \varepsilon$$

for all points ω. In fact, it is again the minimum over all lengths n with this property. (The variables $\tilde{X}$ and X may differ wildly outside the set A, however).

We now want to estimate the average over m terms. We assume $m \ge N_\varepsilon$. For each point $\omega \in \Omega$, we recursively define the sequence

$$n_0(\omega) := 0, \quad n_{j+1}(\omega) := n_j(\omega) + \tilde{n}(T^{n_j}\omega).$$

This sequence increases, so at some point $J = J(\omega)$, it will exceed the number of terms, $n_{J+1}(\omega) \ge m$. Since the step length is bounded, we have $m \le n_J + N_\varepsilon$. The idea was to split the average into sections whose average is close to the limes superior. We obtain the estimate

$$\begin{aligned} \frac{1}{m} \sum_{k=0}^{m-1} \tilde{X}(T^k\omega) &= \frac{1}{m} \left[\sum_{k=n_0(\omega)}^{n_1(\omega)-1} \tilde{X}(T^k\omega) + \sum_{k=n_1(\omega)}^{n_2(\omega)-1} \tilde{X}(T^k\omega) + \ldots + \sum_{k=n_J(\omega)}^{m-1} \tilde{X}(T^k\omega) \right] \\ &\ge \frac{1}{m} \left[(n_1(\omega) - n_0(\omega))(\overline{X}^C(\omega) - \varepsilon) + \ldots + (n_J(\omega) - n_{J-1}(\omega))(\overline{X}^C(T^{n_{J-1}(\omega)}\omega) - \varepsilon) \right] \\ &\ge \frac{n_J(\omega)}{m} (\overline{X}^C(\omega) - \varepsilon) \ge \overline{X}^C(\omega) - \varepsilon - \frac{N_\varepsilon}{m} C. \end{aligned}$$

In the last line, we have used that the variable $\overline{X}^C$ is also invariant under T. We now integrate over the point ω and use the fact that the transformation T is measure-preserving to obtain

$$\int \tilde{X}(\omega)\, dP = \int \left[\frac{1}{m} \sum_{k=0}^{m-1} \tilde{X}(T^k\omega) \right] dP \ge \int \overline{X}^C\, dP - \varepsilon - \frac{N_\varepsilon}{m} C.$$

The left-hand side no longer depends on m, and we can immediately take the limit $m \to \infty$. The integrals of X and $\tilde{X}$ are related by

$$\int X \, dP \geq \int_A X \, dP \geq \int \tilde{X} \, dP - C\varepsilon \geq \int \overline{X}^C \, dP - C\varepsilon - \varepsilon.$$

Taking the limit $\varepsilon \to 0$ and then the limit $C \to \infty$ proves the desired inequality. $\square$

A Measure Theory

A.1 σ-algebras

Definition. Let Ω be a set. An **algebra** is a collection of subsets $\mathcal{A} \subset 2^\Omega$ such that

1. $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$.
2. *Complements.* $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$.
3. *Finite intersections.* $A \in \mathcal{A}$ and $B \in \mathcal{A} \implies (A \cap B) \in \mathcal{A}$.
4. *Finite union.* $A \in \mathcal{A}$ and $B \in \mathcal{A} \implies (A \cup B) \in \mathcal{A}$.

An algebra is said to be a **σ-algebra** if it satisfies the stronger requirement

4'. *Countable union.* If $(A_n)_{n=1}^\infty$ is a countable sequence of subsets in $\mathcal{A}$, then their union is also in the collection, $\bigcup_{n=1}^\infty A_n \in \mathcal{A}$.

In the latter case, the pair $(\Omega, \mathcal{A})$ is also called a **measurable space**.

Remark. By De Morgan's law, $A \cap B = (A^c \cup B^c)^c$, the third condition is actually redundant.

Remark. The intersection of σ-algebras is again a σ-algebra. If $\mathcal{B} \subset 2^\Omega$ is any collection of sets, then the **σ-algebra generated** by this collection, $\sigma(\mathcal{B})$, is the unique smallest σ-algebra that contains this collection. It is given by the intersection $\sigma(\mathcal{B}) = \bigcap_{\mathcal{A}\ \sigma\text{-algebra}, \mathcal{B}\subset\mathcal{A}} \mathcal{A}$.

Example. If $\Omega = X$ is a topological space, then the **Borel σ-algebra** on this space is the σ-algebra generated by the open subsets.

Notation. The measurable space of the real numbers together with the Borel σ-algebra is usually denoted by $(\mathbb{R}, \mathcal{B})$.

Example. Consider the euclidean space $\mathbb{R}^n$. The collection of sets which can be represented as finite, disjoint unions of half-open boxes,

$$\mathcal{A} = \left\{ \biguplus_{k=1}^{N} [a_k, b_k) : N \in \mathbb{N}, a_k, b_k \in (\mathbb{R} \cup \{-\infty, +\infty\})^n \right\},$$

is an algebra. The σ-algebra generated by this collection is precisely the Borel σ-algebra of the space $\mathbb{R}^n$.

Definition. Let $(E_j, \mathcal{A}_j)_{j\in J}$ be a family of measurable spaces. It may well be infinite. On the product space $\prod_{j\in J} E_j$, we define the **product σ-algebra** as the σ-algebra generated by one-dimensional cylinders,

$$\bigotimes_{j\in J} \mathcal{A}_j := \sigma\left(\left\{\pi_k^{-1}(B) : B \in \mathcal{A}_k \text{ for some } k \in J\right\}\right),$$

where $\pi_k : \prod_{j\in J} E_j \to E_k$ is the projection onto the k-the coordinate.

For instance, consider the case where all spaces are equal, $(E_j, \mathcal{A}_j) \equiv (E, \mathcal{A})$, and the family is countably infinite, $J = \mathbb{N}$. Then, the product space is also called the **sequence space** and denoted by $(E^{\mathbb{N}}, \mathcal{A}^{\otimes \mathbb{N}})$. The σ-algebra is generated by sets of the form

$$\pi_k^{-1}(B) = \underbrace{E \times E \times \ldots \times E}_{k-1 \text{ times}} \times B \times E \times \ldots \subset E^{\mathbb{N}}.$$

Example. The Borel σ-algebra of the euclidean space $\mathbb{R}^n$ is equal to the product σ-algebra $\mathcal{B}^{\otimes n}$, where $(\mathbb{R}, \mathcal{B})$ is the space of real numbers with the Borel σ-algebra.

Remark. Given a collection of sets $\mathcal{B}$ that contains another collection $\mathcal{A} \subset \mathcal{B}$, one often wants to show that it also contains the σ-algebra generated by latter, that is $\sigma(\mathcal{A}) \subset \mathcal{B}$. Sometimes, this is not straightforward. The following notions can help:

Definition. A collection of subsets $\mathcal{A} \subset 2^\Omega$ is called a **π-system** if it is non-empty and closed under finite intersections, that is $A \in \mathcal{A}$ and $B \in \mathcal{A} \implies (A \cap B) \in \mathcal{A}$.

Definition. A collection of subsets $\mathcal{A} \subset 2^\Omega$ is called a **Dynkin system** if it satisfies the following conditions:

1. $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$.
2. *Complements.* $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
3. *Disjoint countable union.* If $(A_n)_{n=1}^\infty$ is a countable sequence of *disjoint* subsets in $\mathcal{A}$, $A_i \cap A_j = \emptyset$ for any $i \neq j$, then their union is also in the collection, $\biguplus_{n=1}^\infty A_n \in \mathcal{A}$.

Note that there is no mention of intersections.

It is straightforward to show that every Dynkin system which is also a π-system is already a σ-algebra. However, even more is true:

Proposition A.1 (**Dynkin's theorem**). *Let $\mathcal{B}$ be a Dynkin system and $\mathcal{A}$ be a π-system. If the former contains the latter, $\mathcal{B} \supset \mathcal{A}$, then it also contains the σ-algebra generated by the latter, $\mathcal{B} \supset \sigma(\mathcal{A})$.*

Proof. We want to find a smaller Dynkin system that is closed under intersections. First, consider the collection of sets that have a "good" intersection with sets from the collection $\mathcal{A}$:

$$\mathcal{D}_1 := \{B \in \mathcal{B} : (A \cap B) \in \mathcal{B} \text{ for all } A \in \mathcal{A}\}.$$

Apparently, we have $\mathcal{A} \subset \mathcal{D}_1 \subset \mathcal{B}$. But we can check that this is a Dynkin system: For all $A \in \mathcal{A}$, we can use the fact that the collection $\mathcal{B}$ is a Dynkin system to conclude that

1. $A \cap \emptyset = \emptyset$ and $A \cap \Omega = A$, so both are in the collection $\mathcal{B}$.
2. If $(A \cap B) \in \mathcal{B}$, then $A \cap B^c = (A^c \cup B)^c = (A^c \uplus (A \cap B))^c$ is also in $\mathcal{B}$.
3. If the sets $B_k \in \mathcal{D}_1$ are disjoint, then the set $A \cap (\biguplus_{k=1}^\infty B_k) = \biguplus_{k=1}^\infty (A \cap B_k)$ is also in $\mathcal{B}$.

In a second step, we consider the collection of sets that have "very good" intersections with all sets from this collection, i.e.

$$\mathcal{D}_2 := \{B \in \mathcal{B} : (B \cap A) \in \mathcal{D}_1 \text{ for all } A \in \mathcal{D}_1\}.$$

By construction of the first collection $\mathcal{D}_1$, we have $\mathcal{A} \subset \mathcal{D}_2 \subset \mathcal{D}_1$. But since $\mathcal{D}_1$ is a Dynkin system, we can essentially repeat the same calculations as above and conclude that $\mathcal{D}_2$ is a Dynkin system as well. However, it is also stable under intersections, because if $B, C \in \mathcal{D}_2$, then we have

$$(B \cap C) \cap A = B \cap \underbrace{(C \cap A)}_{\in \mathcal{D}_1} \in \mathcal{D}_1 \quad \text{for all } A \in \mathcal{D}_1,$$

which implies $B \cap C \in \mathcal{D}_2$. In other words, this collection is in fact a σ-algebra, and we have $\sigma(\mathcal{A}) \subset \mathcal{D}_2 \subset \mathcal{B}$ as desired. □

A.2 Measures

Definition. Let $\mathcal{A}$ be an algebra. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a **content** if it is **finitely additive**, i.e. if it satisfies $\mu(\emptyset) = 0$ and

$$\mu(A \cup B) = \mu(A) + \mu(B) \quad \text{whenever } A, B \in \mathcal{A} \text{ are disjoint, } A \cap B = \emptyset.$$

A content is called a **pre-measure** if it is σ-additive, that is if additivity extends to countably infinite unions

$$\mu(A) = \sum_{k=1}^{\infty} \mu(A_k) \quad \text{whenever } A_k \in \mathcal{A} \text{ are disjoint, } A = \biguplus_{k=1}^{\infty} A_k, \text{ and } A \in \mathcal{A}.$$

Note that a pre-measure is defined on an algebra, so the countable union of disjoint sets $A_k \in \mathcal{A}$ does not have to be in the algebra $\mathcal{A}$ again. In this case, the condition does not say anything.

A **measure** $\mu : \mathcal{A} \to [0, \infty]$ is a pre-measure where the underlying algebra $\mathcal{A}$ is also a σ-algebra.

Notation. Let $(B_n)_{n=1}^{\infty}$ be a sequence of sets. We write $B_n \uparrow B$ if the sequence is monotonically increasing, $B_1 \subset B_2 \subset \ldots$, and $B = \bigcup_{n=1}^{\infty} B_n$. Similarly, we write $B_n \downarrow B$ if the sequence is monotonically decreasing, $B_1 \supset B_2 \supset \ldots$, and $B = \bigcap_{n=1}^{\infty} B_n$.

Lemma A.2 (**Continuity** and **pre-measures**)**.** *Let μ be content on an algebra $\mathcal{A}$. If this content satisfies $\mu(\Omega) < \infty$, then the following are equivalent:*

1. *μ is a pre-measure.*
2. *For every sequence of sets $A_n \in \mathcal{A}$ with $A_n \downarrow \emptyset$, we have $\mu(A_n) \to 0$.*

Definition. Consider a σ-algebra $\mathcal{A}$ on a space Ω. A measure μ is called **finite** if $\mu(\Omega) < \infty$. A measure μ is called **σ-finite** if the space can be exhausted by sets with finite content, i.e. there exists a sequence of sets $B_n \in \mathcal{A}$ with $B_n \uparrow \Omega$ and $\mu(B_n) < \infty$.

When the measurable space is also a topological space, we may ask how well a measure can be approximated by knowing only the measures of open sets. The following notion is particularly useful:

Definition. Let X be a topological space and $\mathcal{B}$ the σ-algebra of Borel sets. Then, a measure μ is called **inner regular** if it can be approximated from below by compact sets, that is

$$\mu(A) = \sup\{\mu(K) : K \subset A, K \text{ compact}\} \quad \text{for all Borel sets } A \in \mathcal{B}. \tag{21}$$

Lemma A.3. *Every finite measure on a euclidean space $\mathbb{R}^d$ is inner regular.*

Proof. Consider the collection $\mathcal{C}$ of measurable subsets such that for any $A \in \mathcal{C}$, both A and its complement A^c satisfy the inner regularity property (21). We want to show that it is a Dynkin system.

1. Clearly, $A = \emptyset$ satisfies this property. By exhausting the euclidean space with large cubes, we see that it also holds for $A = \mathbb{R}^d$. The finiteness of the measure is used here.

2. If $A \in \mathcal{C}$, then $A^c \in \mathcal{C}$ by definition.

3. Let $A_n \in \mathcal{C}$ be a countable sequence of disjoint subsets. Let $A = \biguplus_{n=1}^{\infty} A_n$ be the union.

First, we show that $A \in \mathcal{C}$. For each $\varepsilon > 0$, we can find compact sets $K_n \subset A_n$ such that $\mu(K_n) \geq \mu(A_n) - \varepsilon 2^{-n}$. Since the measure is finite and countably additive, there must be an index N such that $\mu(A) \geq \mu(\biguplus_{n=1}^{N} A_n) - \varepsilon$. But the union $K := \biguplus_{n=1}^{N} K_n$ is a compact set and satisfies $\mu(K) \geq \mu(A) - 2\varepsilon$.

Second, we show that $A^c \in \mathcal{C}$. For $\varepsilon > 0$, let $K_n \subset A_n^c$ be compact subsets such that $\mu(K_n) \geq \mu(A_n) - 2^{-n}\varepsilon$.

$$\begin{aligned}\mu(A^c) &= \mu\left(A_1 \cap A_2 \cap \dots\right) \\ &= \mu((A_1 \setminus K_1) \cap \dots) + \mu(K_1 \cap (A_2 \setminus K_2) \cap \dots) + \dots + \mu(K_1 \cap K_2 \cap \dots) \\ &\leq \mu(K_1 \cap K_2 \cap \dots) + \varepsilon.\end{aligned}$$

An infinite intersection of closed and bounded sets is again closed and bounded, so the intersection $K_1 \cap K_2 \dots$ is compact.

Now, we have already mentioned that the disjoint unions of half-open intervals form an algebra. Since an individual half-open interval $[a, b)$ can be written as a countable union of compact intervals, this means that this algebra is a subset of the Dynkin system $\mathcal{C}$. By Dynkin's theorem A.1, this Dynkin system contains the σ-algebra generated by the half-open intervals, which is the Borel σ-algebra. □

A.3 Extension Theorems

Proposition A.4 (**Hahn-Kolmogorov Extension Theorem**). *Let $\mathcal{A}$ be an algebra of subsets and $\mu : \mathcal{A} \to [0, \infty]$ a pre-measure on that algebra. Then, there exists a measure $\tilde{\mu} : \sigma(\mathcal{A}) \to [0, \infty]$ on the generated σ-algebra that extends this pre-measure, $\tilde{\mu}|_{\mathcal{A}} = \mu$. If the pre-measure is σ-finite, then this extension is unique.*

Example. Consider the set of real numbers $\mathbb{R}$. We have already seen that the collection of all disjoint unions of half-open intervals $\biguplus_{k=1}^{n}[a_k, b_k)$ forms an algebra. A compactness argument shows that the sum of lengths is a pre-measure. Applying the extension theorem gives the Lebesgue measure on the real numbers.

Lemma A.5. *Let $\mathcal{A}$ be an algebra and μ be a measure on the generated σ-algebra $\sigma(\mathcal{A})$. Then, for each $\varepsilon > 0$ and each measurable set $A \in \sigma(\mathcal{A})$, there exists a set $B \in \mathcal{A}$ from the algebra such that $\mu(A \triangle B) < \varepsilon$.*

References

[1] S. Varadhan, *Limit theorems*. (2002). (lecture notes)

[2] W. König, *Wahrscheinlichkeitstheorie*. (2013). (lecture notes)