Chapter 1

Events and Probabilities

1. Elementary and semi-elementary probability theory

The standard model of probability theory is the triplet $(\Omega, \mathcal{F}, P)$, where $\Omega$ is a set, $\mathcal{F} \subset 2^\Omega$, and $P$ is a function $\mathcal{F} \to [0,1]$. The set $\Omega$ is called sometimes a sample space or a probability space. Sets $A \in \mathcal{F}$ are called events, and the number $P(A)$ is called probability of event $A$. We say that event $A$ occurred, if $\omega \in A$, and we interpret $P(A)$ as the “likelihood” that event $A$ occurred.

Notation. We will follow notation from [Billingsley]

- $\Omega$ - an abstract set representing the sample space of some experiment. Elements of $\Omega$ will be denoted by $\omega$.
  - Examples are finite sets, intervals, $\Omega = [0,1]$, subsets of $\mathbb{R}^n$, $\mathbb{R}^\infty$, $C[0,1]$ - these are all complete and separable metric spaces.
- $2^\Omega$ - the set of all subsets of $\Omega$. We use upper case letters like $A, B, C, \ldots$ for the subsets - some (but not all) will be interpreted as "events", that is collections of simple events (points of $\Omega$).
  - For example, if $\Omega = \mathbb{R}^\infty$, one might be interested in the ”events"
    $A = \left\{ \omega = (x_k)_{k \in \mathbb{N}} \in \mathbb{R}^\infty : \lim_{n \to \infty} \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = 17 \right\}$
    $B = \left\{ \omega = (x_k)_{k \in \mathbb{N}} \in \mathbb{R}^\infty : \sup_k x_k \leq 17 \right\}$
- The empty set is $\emptyset$ - in handwriting this needs to be carefully distinguished from the Greek letters $\varnothing$ or $\Phi$.
- Collections of sets will be denoted by scripted letters like $A$ or $\mathcal{F}$.
- For union and intersections we use $\bigcup_{t \in T} A_t = \{ \omega \in \Omega : \exists_{t \in T}, \omega \in A_t \}$ and $\bigcap_{t \in T} B_t = \{ \omega \in \Omega : \forall_{t \in T}, \omega \in A_t \}$
It is natural to expect that the events form a field.

**Definition 1.1.** A class $\mathcal{F}$ of subsets of $\Omega$ is a field if:

(i) $\Omega \in \mathcal{F}$

(ii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

(iii) if $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$

By induction, if $A_1, A_2, \ldots, A_n \in \mathcal{F}$ then $A_1 \cup \cdots \cup A_n \in \mathcal{F}$. By DeMorgan’s law (Theorem A.3), a field is also closed under intersections, $\bigcap_{1 \leq j \leq n} A_j \in \mathcal{F}$. In particular, we can replace axiom (iii) by

(iii’) if $A, B \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$

For a given $\Omega$, the smallest field is $\{\Omega, \emptyset\}$. The largest field is the so-called power set $2^\Omega$ - the set of all subsets of $\Omega$.

**Proposition 1.1.** Given a collection of sets $\mathcal{A}$, there exists a unique field $\mathcal{F}$ such that $\mathcal{A} \subset \mathcal{F}$ and if $\mathcal{G}$ is a field that contains $\mathcal{A}$ then $\mathcal{F} \subset \mathcal{G}$. Then $\mathcal{F}$ is called the field generated by $\mathcal{A}$.

**Proof.** This is a "standard construction":

$$\mathcal{F} = \bigcap \{\mathcal{G} : \mathcal{G} \text{ is a field, and } \mathcal{A} \subset \mathcal{G}\}$$

The following example describes a version of a Borel field.

**Example 1.1.** The class $\mathcal{B}_0$ of finite unions of disjoint left-open right-closed subintervals of $(0, 1]$, is a field.

**Proof.** $(0, 1] \in \mathcal{B}_0$. If $A = \bigcup_{j=1}^{K} (a_j, b_j]$ with $a_1 < b_1 \leq a_2 < b_2 \cdots \leq a_K < b_K$ then $A^c = (0, a_1] \cup (b_1, a_2] \cup \cdots \cup (b_{K-1}, a_K] \cup (b_K, 1]$, where some of the intervals might be empty.

If $A = \bigcup_j I_j$ and $B = \bigcup_k J_k$ then $A \cup B = \bigcup_{j,k} I_j \cap J_k$ and intersections $I_j \cap J_k$ are disjoint, possibly empty, intervals of the form $(a, b]$.

Similarly, for $\Omega := (0, 1] \times (0, 1]$, the set of finite unions of rectangles $(a, b] \times (c, d] \subset$ is a field.

**Remark 1.2.** The field $\mathcal{B}_0$ in Example 1.1 is the smallest field of subsets of $(0, 1]$ that contains intervals $(a, b]$.

**Question 1.1.** What is the smallest field of subsets of $\mathbb{R}$ that contains all half-lines $(-\infty, b]$?

### 1.1. Finitely additive probabilities. We want to assign the number $P(A)$, as a “measure” of the likelihood that the event $A$ occurred.

**Definition 1.2.** Let $\mathcal{F}$ be a field. A function $P : \mathcal{F} \to \mathbb{R}$ is a finitely-additive probability measure if it satisfies the following conditions

(i) $0 \leq P(A) \leq 1$ for $A \in \mathcal{F}$

(ii) $P(\emptyset) = 0$, $P(\Omega) = 1$.

(iii) If $A, B \in \mathcal{F}$ are disjoint, then $P(A \cup B) = P(A) + P(B)$. 
A function $P : \mathcal{F} \to \mathbb{R}$ is a probability measure on $\mathcal{F}$ if it is finitely-additive and satisfies the following continuity condition.

(iv) If $A_1 \supset A_2 \supset \ldots$ are sets in $\mathcal{F}$ and $\bigcap_k A_k = \emptyset$, then $\lim_{n \to \infty} P(A_n) = 0$

**Proposition 1.3.** Suppose $P$ is a probability measure on the field $\mathcal{F}$ of subsets of $\Omega$. For $A, B \in \mathcal{F}$ we have

(i) $B \subset A$ implies $P(A) \leq P(B)$

(ii) $P(A^c) = 1 - P(A)$

(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iv) If $A_1, A_2, \ldots \in \mathcal{F}$ are such that $\bigcup_{k=1}^\infty A_k \in \mathcal{F}$ then

\[ P\left( \bigcup_{k=1}^\infty A_k \right) \leq \sum_{k=1}^\infty P(A_k) \]

**Proof.** (i) follows from $A = B \cup (A \setminus B)$ by finite additivity. (The proof gives $P(B \setminus A) = P(B) - P(A)$)

(ii) follows from $\Omega = A \cup A^c$

(iii) is a special case of Exercise 1.1. For example, using formula from the proof of (1): $P((A \cup B) \setminus A) = P(A \cup B) - P(A)$. Since $(A \cup B) \setminus A = (A \cup B) \cap A^c = B \cap A^c = B \setminus (A \cap B)$ using the formula again, $P((A \cup B) \setminus A) = P(B) - P(A \cap B)$. So $P(A \cup B) - P(A) = P(B) - P(A \cap B)$ and the formula follows. (There are numerous other proofs!)

(iv) Let $B = \bigcup_{k=1}^\infty A_k$ and $B_n = \bigcup_{k=1}^n A_k$. Then (mathematical induction!) $P(B_n) \leq \sum_{k=1}^n P(A_k)$ and

\[ P(B) = P(B \setminus B_n) + P(B_n) \]

Events $B \setminus B_n$ are decreasing and $\bigcap_{n=1}^\infty (B \setminus B_n) = B \cap (\bigcup_{n=1}^\infty B_n)^c = \emptyset$ so as $n \to \infty$, we have $P(B \setminus B_n) \to 0$ by continuity axiom.

**Remark 1.4.** If $A_k$ are pairwise disjoint and $\bigcup_{k=1}^\infty A_k \in \mathcal{F}$ then then the above proof gives countable additivity:

(iii+) If $A_1, A_2, \ldots \in \mathcal{F}$ are pairwise disjoint and $\bigcup_{k=1}^\infty A_k \in \mathcal{F}$, then $P(\bigcup_{k=1}^\infty A_k) = \sum_{k=1}^\infty P(A_k)$.

Indeed, $\sum_{k=1}^n P(A_k) \leq P(B) \leq P(B \setminus B_n) + \sum_{k=1}^n P(A_k)$ and we can pass to the limit as $n \to \infty$.

So axioms (iii) and (iv) are often combined together into countable additivity (iii+)

Other equivalent versions of continuity or countable additivity are:

(iv') If $A_1 \subset A_2 \subset \ldots$ are sets in $\mathcal{F}$ and $\bigcup_{k=1}^\infty A_k \in \mathcal{F}$, then $P(\bigcup_{k=1}^\infty A_k) = \lim_{n \to \infty} P(A_n)$

(iv'') If $A_1 \supset A_2 \supset \ldots$ are sets in $\mathcal{F}$ and $\bigcap_{k=1}^\infty A_k \in \mathcal{F}$, then $P(\bigcap_{k=1}^\infty A_k) = \lim_{n \to \infty} P(A_n)$

(Recall Theorem A.3.)

Constructions of finitely additive continuous measures are somewhat more involved.

1.1.1. Example: Lebesgue measure on the unit interval. In this example we consider $\Omega = (0, 1]$ and the field $\mathcal{B}_0$ from Example 1.1.

For $A = \bigcup_{k=1}^n I_k \in \mathcal{B}_0$ with disjoint $I_k$, define $\lambda(A) = \sum_{k=1}^n |I_k|$.

**Theorem 1.5.** $\lambda$ is a well defined (continuous) probability measure on the field $\mathcal{B}_0$. 

1. Events and Probabilities

Theorem 1.6 (Lebesgue). If $P$ is a (continuous) probability measure on $\mathcal{B}_0$ then the function $F(x) := P((-\infty, x])$ has the following properties:

(i) $F$ is non-decreasing
(ii) $\lim_{x \to -\infty} F(x) = 0$
(iii) $\lim_{x \to \infty} F(x) = 1$
(iv) $F$ is right-continuous, i.e., $\lim_{y \uparrow x} F(y) = F(x)$

Conversely, if $F$ is a function with properties (i)-(iv) then there exists a (unique) continuous probability measure $P$ on $\mathcal{B}_0$ such that $F(x) := P((-\infty, x])$ for all $x \in \mathbb{R}$.

Sketch of the proof. The proof of (iv) may require some care: For rational $r \downarrow x$, we have $(0, x] = \bigcap_{r > x} (0, r]$. Any real $y$ lies between two rational numbers.

The proof of converse has two parts. For $A \in \mathcal{B}_0$ given by $A = \bigcup_{j=1}^n (a_j, b_j]$, the definition $P(A) = \sum_{j=1}^n (F(b_j) - F(a_j))$ does not depend on the representation. (Here, we set $F(-\infty) = 0$ and $F(\infty) = 1$.) Uniqueness is an obvious consequence of finite additivity.

Then we need to verify continuity. This proof can proceed similarly to the proof of Theorem 1.5. (See hints for Exercise 1.11.)

A generalization of the above construction is based on the concept of a semi-algebra.

Definition 1.3. A collection $\mathcal{S}$ of subsets of $\Omega$ is called semi-algebra, or a semi-ring, if

(i) $\emptyset \in \mathcal{S}$
(ii) $\mathcal{S}$ is closed under intersections, i.e. if $A, B \in \mathcal{S}$ then $A \cap B \in \mathcal{S}$
(iii) If $A, B \in \mathcal{S}$ then $B \setminus A$ is a finite union of sets in $\mathcal{S}$.

The main (motivating) example of a semi-algebra is the family of rectangles in $\mathbb{R}^2$, and more generally, in $\mathbb{R}^d$. 

(Omitted in 2019)

Proof. Since the representation $A = \sum_{k=1}^n I_k \in \mathcal{B}_0$ is not unique, we need to make sure that $\lambda$ is well defined. Write $A = \bigcup_{k} I_k = \bigcup_{j} J_j$ as the finite sums of disjoint intervals. Then $I_k = I_k \cap A = \bigcup_{j} I_k \cap I_j$ so by finite additivity $\sum_k |I_k| = \sum_k \sum_j |I_k \cap I_j|$ and similarly $\sum_j |J_j| = \sum_j \sum_k |I_k \cap I_j|$. This shows that $\sum_j |J_j| = \sum_k |I_k|$, so $\lambda(A)$ is indeed well defined.

To prove continuity, we proceed by contrapositive. Suppose that $A_n \supset A_{n+1}$ are such that $\lambda(A_n) > \delta > 0$. Choose $B_n \subset K_n \subset A_n$ such that $\lambda(A_n) - \lambda(B_n) < 2^{-n}$ and $K_n$ is compact. Then $P(A_n) - P(B_1 \cap \cdots \cap B_n) = P(\bigcup_{k=1}^n A_n \setminus B_k) \leq P(\bigcup_{k=1}^n A_k \setminus B_k) \leq \sum_{k=1}^\infty 2^{-k} = \delta/2$. So $P(B_1 \cap \cdots \cap B_n) \geq \delta/2 > 0$, and $K_1 \cap \cdots \cap K_n \neq \emptyset$. Thus, $\bigcap_{n=1}^\infty A_n \supset \bigcap_{n=1}^\infty K_n \neq \emptyset$. 

□

Another argument for countable additivity appears in [Billingsley, Theorem 1.3].
The following is a version of [Durrett, Theorem 1.1.4] adapted to probability measures.

**Theorem 1.7.** Let $S$ be a semi-algebra. Suppose that $P : S \to [0, 1]$ is additive, countably subadditive, i.e., if $A = \bigcup_{n=1}^{\infty} A_n$ is in $S$ for pairwise disjoint sets $A_n \in S$ then $P(A) \leq \sum_{n=1}^{\infty} P(A_n)$. If $P(\emptyset) = 0$, then $P$ has a unique extension onto the field $F$ generated by $S$, and this extension is continuous.

**Proof.**

**Question 1.2.** Why not to require uncountable continuity of probability measures?

(Compare Exercise 1.4 and Exercise 1.4.)

### 2. Sigma-fields

Given an infinite sequence $\{A_n\}$ of events, it is convenient to allow also more complicated events such as $A^* = \bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k$ that events $A_k$ occur infinitely often. This motivates the following.

**Definition 1.4.** A class $F$ of subsets of $\Omega$ is a $\sigma$-field if it is field and if it is also closed under the formation of countable unions:

(iii+) If $A_1, A_2, \cdots \in F$ then $\bigcup_{n \in \mathbb{N}} A_n \in F$.

By an application of DeMorgan’s law (Theorem A.3), (iii+) can be replaced by

If $A_1, A_2, \cdots \in F$ then $\bigcap_{n \in \mathbb{N}} A_n \in F$.

As for the fields, for each set $\Omega$ there is a smallest $\sigma$-fields $\{\emptyset, \Omega\}$ and the largest $\sigma$-field which is the power set $2^\Omega$. As previously, a collection of sets $A$, generates the $\sigma$-field, the smallest $\sigma$-fields that contain all of the sets in $A$.

**Proposition 1.8.** Suppose $A$ is a collection of subsets of $\Omega$. There exist a unique $\sigma$-field $F$ with the following properties:

(i) $A \in A$ implies $A \in F$. That is, $A \subset F$.

(ii) If $G$ is a $\sigma$-field such that $A \subset G$ then $F \subset G$.

We write $F = \sigma(A)$, and call $F$ the $\sigma$-field generated by $A$.

**Proof.** Uniqueness is a consequence of (ii)

To show that $F$ exists, consider a set $M$ of all sigma-fields $G$ with the property that $A \subset G$. Since $2^\Omega \in M$, this is a nonempty family of sets. Define

$$F = \bigcap_{G \in M} G.$$ 

Then $F$ is a sigma-field with the required properties, because the intersection of sigma-fields is a sigma-field (can you verify this?).

**Definition 1.5.** The Borel $\sigma$-field is the sigma field generated by all open sets of $\Omega$.

In particular, the Borel $\sigma$-field $B_\mathbb{R}$
Proposition 1.9. For \( \Omega = \mathbb{R} \), the Borel \( \sigma \)-field \( \mathcal{B}_{\mathbb{R}} \) is generated by the intervals \( \{(a, b) : a < b\} \).

For \( \Omega = \mathbb{R}^d \), the Borel sigma-field \( \mathcal{B}_{\mathbb{R}^d} \) is generated by the rectangles \( \prod_{k=1}^{d} (a_k, b_k) \).

Sketch for \( \mathbb{R} \). It is clear that \( \mathcal{B}_{\mathbb{R}} \) is generated by open intervals. So \( (a, b) = \bigcap_{n=1}^{\infty} \left( a, b + \frac{1}{n} \right) \in \mathcal{B}_{\mathbb{R}} \).

So the sigma field \( \mathcal{F} \) generated by the intervals \( \{(a, b) : a < b\} \) is in Borel \( \sigma \)-field. Conversely, \( (a, b) = \bigcup_{n=1}^{\infty} \left( a, b - \frac{1}{n} \right) \) so open intervals are in the \( \sigma \)-field \( \mathcal{F} \) generated by the intervals \( \{(a, b) : a < b\} \). Thus \( \mathcal{B}_{\mathbb{R}} \subset \mathcal{F} \). □

Note that Borel-field \( \mathcal{B}_{0} \subset 2^{\mathbb{R}} \) generated by all intervals \( (a, b] \) consist of finite unions of intervals \( (-\infty, b], (a, b], (b, \infty) \).

Remark 1.10. \( \mathcal{B}_{\mathbb{R}^2} \) is also generated by open disks. For example, a unit disk \( x^2 + y^2 \leq 1 \) is the union of the countable family of rectangles \( (-a_n, a_n) \times (-b_n, b_n) \) with rational \( a_n, b_n \) such that \( a_n^2 + b_n^2 \leq 1 \).

2.1. Probability measures.

Definition 1.6. If \( \mathcal{F} \) is a \( \sigma \)-field of subsets of \( \Omega \) and \( P \) is a probability measure on \( \mathcal{F} \), then the triple \( (\Omega, \mathcal{F}, P) \) is called a probability space. The sets \( A \in \mathcal{F} \) are called events.

Here are some examples of probability measures.

Example 1.2 (Discrete measures). Let \( \mathcal{F} = 2^{\Omega} \) and fix \( \omega_0 \in \Omega \). Then \( P(A) = I_A(\omega_0) \) is a probability measure, sometimes called the point mass, denoted by \( \delta_{\omega_0} \). Formally,

\[
\delta_{\omega}(A) = \begin{cases} 
1 & \omega \in A \\
0 & \omega \not\in A.
\end{cases}
\]

Since a convex combination of probability measures is a probability measure, additional examples of a probability measures on \( (\Omega, \mathcal{F}) \) arise by choosing a finite or infinite sequence \( \omega_k \in \Omega \) and a sequence of positive numbers \( p_k \), see Example 1.4.

Example 1.3 (Discrete probability space). Let \( \mathcal{F} \) be the \( \sigma \)-field of all subsets of a countable \( \Omega = \{\omega_1, \omega_2, \ldots\} \) Suppose \( p_1, p_2, \ldots \) is a sequence of nonnegative numbers such that \( \sum_{k=1}^{\infty} p_k = 1 \). Define

\[
P(A) = \sum_{k: \omega_k \in A} p_k
\]

Then \( P \) is a probability measure.

Proof. This is not entirely trivial. If \( A = \bigcup_j A_j \) with disjoint sets then \( P(A) = \sum_{\omega_i \in \bigcup_j A_j} p_i \) does not depend on the order of summation, and equals to the iterated series \( \sum_{j=1}^{\infty} \sum_{i: \omega_i \in A_j} p_i \). □

Example 1.4 (Discrete probability measure). Let \( \mathcal{F} \) be the \( \sigma \)-field of all subsets of an infinite set \( \Omega \). Suppose \( \omega_1, \omega_2, \ldots \in \Omega \) are fixed, and \( p_1, p_2, \ldots \) is a sequence of nonnegative numbers such that \( \sum_{k=1}^{\infty} p_k = 1 \). Define

\[
P(A) = \sum_{k: \omega_k \in A} p_k
\]

Then \( P \) is a probability measure.
Proof. The example can be written using the point masses introduced in Example 1.2. That is, we write \( P = \sum_{k=1}^{\infty} p_k \delta_k \). Let's verify that the axioms of probability indeed hold true.

(i) Clearly, \( P(\Omega) \) and \( P(A) \geq 0 \).

(ii) If \( A, B \) are disjoint, then considering four possible cases we check that \( \delta_\omega * (A \cup B) = \delta_\omega(A) + \delta_\omega(B) \). Therefore,

\[
P(A \cup B) = \sum_{k=1}^{\infty} p_k (\delta_k(A) + \delta_k(B)) = \sum_{k=1}^{\infty} p_k \delta_k(A) + \sum_{k=1}^{\infty} p_k \delta_k(B)
\]

(iii) To verify continuity of \( P \), fix a sequence \( B_{n+1} \subset B_n \) of subsets of \( \Omega \) such that \( \bigcap B_n = \emptyset \). Fix \( \varepsilon > 0 \) and choose \( N \) such that \( \sum_{k=N}^{\infty} p_k \neq \varepsilon \).

Then for every \( n \) we have \( P(B_n) \leq \sum_{k=1}^{N} \delta_k(B_n) + \varepsilon \). Moreover, since \( \bigcap B_n = \emptyset \) and the sets are decreasing, for every \( k \) there is \( n_k \) such that for \( n > n_k \) we have \( \omega_k \notin B_n \). Therefore, for \( n > M = \max\{n_1, n_2, \ldots, n_N\} \) we have \( \sum_{k=1}^{N} \delta_k(B_n) = 0 \) and therefore for every \( \varepsilon > 0 \) there is \( M \) such that \( P(B_n) < \varepsilon \) for all \( n > M \)

i.e., \( P(B_n) \to 0 \).

\( \square \)

The numbers \( p_k \) are sometimes called the “probability mass function”, or the “probability density function” (as this is the density with respect to the counting measure on points \( \{x_1, x_2, \ldots\} \)).

More generally, if \( P_1, P_2, \ldots \) is a sequence of probability measures on the common field or \( \sigma \)-field \( F \) and \( p_1, p_2, \ldots \) is a sequence of nonnegative numbers such that \( \sum_{k=1}^{\infty} p_k = 1 \), then \( Q(A) = \sum_{k=1}^{\infty} p_k P_k(A) \) is also a probability measure on \( F \). The probabilistic interpretation is that we have a sequence of experiments described by probability measures \( P_k \) and we want to model a new experiment with additional randomization where the \( k \)-th experiment is performed with probability \( p_k \).

In particular, a sequence \( \omega_n \in \Omega \) together with a sequence of nonnegative numbers \( p_n \) such that \( \sum_{n=1}^{\infty} p_n = 1 \) determines probability measure

\[
P = \sum_{n=1}^{\infty} p_n \delta_{\omega_n}.
\]

Measure \( P \) is the measure described in Examples 1.3 and 1.4.

### 2.1.1. Examples of discrete distributions on \( \Omega = \mathbb{R} \) or \( \Omega = \mathbb{N} \).

#### Example 1.5 (Binomial distribution). For fixed integer \( n \) and \( 0 < p < 1 \), take \( \omega_k = k \) with \( k \leq n \) and

\[
p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots n.
\]

The binomial formula shows that \( \sum_{k=0}^{n} p_k = 1 \). Notation: \( Bin(n, p) \)

#### Example 1.6 (Poisson distribution). For \( \lambda > 0 \) take \( \omega_k = k \) with \( k = 0, 1, \ldots \), and

\[
p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots
\]
Notation: \( \text{Poiss}(\lambda) \)

**Example 1.7** (Polya’s distribution). For \( r > 0 \) and \( 0 < p < 1 \) take \( \omega_k = k \) and

\[
p_k = \frac{\Gamma(r+k)}{k!\Gamma(r)} (1-p)^r p^k, \quad k = 0, 1, \ldots
\]

Notation: \( \text{NB}(p,r) \)

---

### Required Exercises

**Exercise 1.1.** Prove the **inclusion-exclusion** formula\(^1\)

\[
P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j) - \sum_{1 \leq j_1 < j_2 \leq n} P(A_{j_1} \cap A_{j_2}) + \sum_{1 \leq j_1 < j_2 < j_3 \leq n} P(A_{j_1} \cap A_{j_2} \cap A_{j_3}) - \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n)
\]

**Solution:** Proceed by induction. Countable additivity is not used in this proof.

**Exercise 1.2** (statistics). Prove **Boole’s inequality**\(^2\)

\[
P(\bigcup_{j=1}^{n} A_j) \leq \sum_{j=1}^{n} P(A_j)
\]

**Solution:** Note that \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B) \). Now proceed by induction, with \( A = \bigcup_{j=1}^{n} A_j \) and \( B = A_{n+1} \) for the induction step.

**Exercise 1.3.** For a probability space \((\Omega, \mathcal{F}, P)\), if \( B_1, B_2, \ldots \) is a sequence of events such that \( \sum_{k=1}^{n} P(B_k) > n-1 \), show that \( P(\bigcap_{k=1}^{n} B_k) > 0 \).

**Solution:** Proceed by contrapositive. Suppose \( 1 = P(\bigcup_{k=1}^{n} B_k) \leq \sum_{k=1}^{n} P(B_k) = n - \sum_{k=1}^{n} P(B_k) \). Thus \( \sum_{k=1}^{n} P(B_k) \leq n - 1 \).

**Exercise 1.4.** For a probability space \((\Omega, \mathcal{F}, P)\), suppose \( \{B_n : n \in \mathbb{N}\} \) are events with \( P(B_n) = 1 \). Show that

\[
P \left( \bigcap_{n=1}^{\infty} B_n \right) = 1
\]

**Solution:** \( 1 \geq P(\bigcap_{n=1}^{\infty} B_n) = 1 - P(\bigcup_{n=1}^{\infty} B_n^c) \geq 1 - \sum_{n=1}^{\infty} P(B_n^c) = 1 \). In other words, we prove that the complement of \( \bigcap_{n=1}^{\infty} B_n \) has zero probability.

**Exercise 1.5.** Without using Proposition 1.9, show that open intervals \((a,b)\) and closed intervals \([a,b]\) are in the sigma-field generated by the intervals \((a,b)\) in \( \mathbb{R} \).

---

\(^1\)This can also be written as

\[
\Pr \left( \bigcup_{k=1}^{n} A_k \right) = \sum_{M=1}^{n} (-1)^{M-1} \sum_{1 \leq j_1 < j_2 < \cdots < j_M \leq n} \Pr(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_M})
\]

\(^2\)Named for G. Boole (1815-1864). Used for **Bonferroni’s correction** in multiple hypothesis testing.
Solution: Clearly \([a, b] = \bigcap_{n=1}^{\infty} (a - 1/n, b]\). Let \(N\) be an integer such that \(N(b - a) > 1\). Then \((a, b) = \bigcup_{n=N}^{\infty} (a, b - 1/n]\).

Additional Exercises

**Exercise 1.6.** Without using Proposition 1.9, show that the open triangle \(T = \{(x, y) : x > 0, y > 0, x + y < 1\}\) is in the sigma-field generated by the rectangles \((a, b] \times (c, d]\) in \(\mathbb{R}^2\).

**Exercise 1.7.** Suppose that \(F_n\) are fields satisfying \(F_n \subseteq F_{n+1}\). Show that \(\bigcup_{n} F_n\) is a field.

**Exercise 1.8.** Suppose \(P\) is a finitely additive measure on a field \(F\). Show that if \(A_1, \ldots, A_n, \ldots\) are disjoint then the series \(\sum_{n=1}^{\infty} P(A_n)\) converges.

**Solution:** For additive measures, \(\sum_{j=1}^{n} P(A_j) = P(\bigcup_{j=1}^{n} A_j) \leq 1\). Thus the sequence of partial sums is increasing and bounded, so it converges to a number in \([0, 1]\).

**Exercise 1.9.** Prove that continuous finitely-additive probability measure on a field is countably additive. That is, show that property (iii+) of Remark 1.4 follows from the axioms (i)-(iv) of Definition 1.2.

**Exercise 1.10.** If \(P_1, P_2, \ldots,\) is a sequence of continuous probability measures on the field \(F\) and \(p_1, p_2, \ldots\) is a sequence of nonnegative numbers such that \(\sum_{k=1}^{\infty} p_k = 1\), show that \(Q(A) = \sum_{k=1}^{\infty} p_k P_k(A)\) is also continuous.

**Exercise 1.11.** Suppose \(\Omega\) is a metric space and \(F\) is a field of subsets of \(\Omega\). Suppose that \(P\) is a finitely additive probability measure on \(F\).

Let's say that \(P\) is a tight probability measure if for every \(A \in F\) with \(P(A) > 0\) and \(\varepsilon > 0\) there exist \(B \in F\) and a compact set \(K\) such that \(B \subseteq K \subseteq A\) and \(P(A) < P(B) + \varepsilon\).

(i) In the setting of Theorem 1.5, show that the Lebesgue measure on \(B_0\) is tight.

(ii) Show that a tight finitely additive probability measure is countably-additive.

**Hint:** Proceed by contrapositive!\(^3\)

**Solution:** The solution is on page 17.

**Exercise 1.12.** Let \(\Omega\) be an infinite set. Consider the following classes of subsets of \(\Omega:\)

\[ F_n = \{ A \subseteq \Omega : \text{A has at most } n \text{ elements or } A^c \text{ has at most } n \text{ elements} \} \]

Then we have the following facts:

Plan of proof: Suppose \(A_1 \supset A_2 \supset \ldots\) are sets in \(F\) such that there exists \(\delta > 0\) with \(P(A_n) > \delta\) for all \(n\). Using tightness, we can find compact sets \(K_1, K_2, \ldots\) and sets \(B_j \in F\) such that \(B_j \subseteq K_j\) and \(B_1 \cap B_2 \cap \cdots \cap B_n\) has positive probability. In fact, we can find such \(B_j\) with \(P(B_1 \cap B_2 \cap \cdots \cap B_n) \geq \delta(1 - \sum_{j=1}^{n} 1/2^j) > 0\).

Since every finite intersection \(K_1 \cap K_2 \cap \cdots \cap K_n\) contains \(B_1 \cap B_2 \cap \cdots \cap B_n\), we see that \(\bigcap_{n} K_n\) is nonempty. So \(\bigcap_{n} A_n\) cannot be empty.
• $F_n \subset F_{n+1}$
• $F_0$ is a $\sigma$-field
• For $n \geq 1$, class $F_n$ is not a field
• $\bigcup_n F_n$ is a field.
• $\bigcup_n F_n$ is not a $\sigma$-field.

Exercise 1.13 (Compare Exercise 1.4). Suppose $\{B_t : t \in T\}$ are events with $P(B_t) = 1$. Give an example where $\bigcap_{t \in T} B_t = \emptyset$ so $P(\bigcap_{t \in T} B_t) = 0$. Hint: Lebesgue measure on Borel $(0,1]$

Exercise 1.14. Let $\Omega$ be a nonempty set and $C$ be the class of one-element sets. Show that if $A \in \sigma(C)$ then either $A$ is countable or $A^c$ is countable.

Exercise 1.15. Suppose $A$ and $B$ are $\sigma$-fields of subsets of $\Omega$. Let $F = A \land B$ be the smallest $\sigma$-field containing both $A$ and $B$. Show that $F$ is generated by sets of the form $A \cap B$ where $A \in A$ and $B \in B$.

Exercise 1.16. The field $F(A)$ generated by a class $A$ of subsets of $\Omega$ is defined as the intersection of all fields in $\Omega$ containing all of the sets in $A$.
• Show that $F(A)$ is indeed a field, that $A \subset F(A)$ and that $F(A)$ is minimal in the sense that if $G$ is a field and $A \subset G$ then $F(A) \subset G$.
• Show that if $A$ is nonempty then $F(A)$ is the class of sets of the form $\bigcup_{j=1}^{m} A_j$ where sets $A_j$ are disjoint and are of the form $B = \bigcap_{i=1}^{n} A_i$ where either $A_i \in A$ or $A_i^c \in A$.

Exercise 1.17. For $\Omega = (0,1]$ and any $A \subset \Omega$ define
$$P^* = \inf \left\{ \sum_k |B_k| : B_k \in B_0, \bigcup_{k=1}^{\infty} B_k \supset A \right\}$$
where $|B|$ is the sum of lengths of intervals forming $B$.
(i) Show that $0 \leq P^*(A) \leq 1$
(ii) Show that $P^*(A \cup B) \leq P^*(A) + P^*(B)$
(iii) Show that $P^*|_{B_0} = \lambda$, the Lebesgue measure from Theorem 1.5.
(iv) Show that $P^*(\{x\}) = 0.$
Chapter 2

Probability measures

The main results about probability measures are the following two facts:

**Theorem 2.1** (extension). If $P$ is a (continuous) probability measure on a field $\mathcal{F}_0$ then it has a unique extension to $\mathcal{F} = \sigma(\mathcal{F}_0)$.

**Theorem 2.2** (uniqueness). Suppose $\mathcal{F} = \sigma(\mathcal{A})$ and $\mathcal{A}$ is closed under finite intersections (the so called $\pi$-system). If two probability measures on $\mathcal{F}$ agree on $\mathcal{A}$ then they are equal also on $\mathcal{F}$.

The details are somewhat technical, and not very illuminating.

1. Existence

**Theorem 2.3** (Caratheodory). A (countably additive) probability measure on a field has an extension to the generated $\sigma$-field

**Proof of Theorem 2.3.** Let $\mathcal{F}_0$ be a field of subsets of $\Omega$ and let $P_0$ be a probability measure on $\mathcal{F}_0$. Put $\mathcal{F} = \sigma(\mathcal{F}_0)$.

For each subset $A$ of $\Omega$, define the outer measure

\[ P^*(A) = \inf \left\{ \sum_{n=1}^{\infty} P_0(A_n) : A_n \in \mathcal{F}_0, \bigcup_{n=1}^{\infty} A_n \supset A \right\} \]

**Question 2.1.** Can $P^*(A) = \infty$?

Let's first check that $P^*$ is a genuine extension of $P_0$ to a set function defines on all subsets of $\Omega$.

**Proposition 2.4.** $P^*$ and $P$ agree on $\mathcal{F}_0$.

**Proof.** (Omitted in 2019) Suppose $A \in \mathcal{F}_0$. Clearly, $P^*(A) \leq P(A)$ as an infimum. Given $\varepsilon > 0$ choose $A_n \in \mathcal{F}_0$ such that $A \subset \bigcup_{n=1}^{\infty} A_n$ and $P^*(A) + \varepsilon > \sum_{n=1}^{\infty} P(A_n)$. Then $A = \bigcup_{n=1}^{\infty} (A_n \cap A)$ and $A \in \mathcal{F}_0$, so by countable subadditivity $P(A) \leq \sum_{n=1}^{\infty} P(A \cap A_n) \leq \sum_{n=1}^{\infty} P(A_n) < P^*(A) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, this shows that indeed $P(A) = P^*(A)$. 

11
In general, $P^*$ is not additive, at least not on $2^\Omega$, but it still has a number of nice properties.

**Proposition 2.5.** The outer probability has the following properties:

(i) $P^*(\emptyset) = 0$;

(ii) $P^*(A) \geq 0$

(iii) $A \subset B$ implies $P^*(A) \leq P^*(B)$

(iv) $P^*(\bigcup_n A_n) \leq \sum_n P^*(A_n)$

**Proof.** (Omitted in 2019) Without loss of generality we may assume $\sum_n P^*(A_n) < \infty$. To prove (4), choose sets $B_{nk} \in \mathcal{F}_0$ such that $A_n \subset \bigcup_k B_{nk}$ and $P^*(A_n) \leq \varepsilon/2^n + \sum_k P_0(B_{nk})$. Then $\bigcup_n A_n \subset \bigcup_{n,k} B_{nk}$ and $P^*(\bigcup_n A_n) \leq \sum_{n,k} P_0(B_{nk}) = \sum_n \sum_k P_0(B_{nk}) \leq \varepsilon + \sum_n P^*(A_n)$.

Next, consider the class $\mathcal{M}$ of subsets $A$ of $\Omega$ with the property that

\begin{equation}
P^*(A \cap E) + P^*(A^c \cap E) = P^*(E) \text{ for all } E \subset \Omega
\end{equation}

Note that by subadditivity of $P^*$, identity (2.2) is equivalent to inequality

\begin{equation}
P^*(A \cap E) + P^*(A^c \cap E) \leq P^*(E) \text{ for all } E \subset \Omega
\end{equation}

(Omitted in 2019)

**Lemma 2.6.** $\mathcal{M}$ is a field.

**Proof.** Clearly, $\Omega \in \mathcal{M}$ and if $A \in \mathcal{M}$ then $A^c \in \mathcal{M}$. It remains to show that if $A, B \in \mathcal{M}$ then $A \cap B \in \mathcal{M}$. Choose arbitrary $E \subset \Omega$.

\[
P^*(E) = P^*(A \cap E) + P^*(A^c \cap E)
\]

\[
= P^*(B \cap A \cap E) + P^*(B^c \cap A \cap E) + P^*(B \cap A^c \cap E) + P^*(B^c \cap A^c \cap E)
\]

\[
\geq P^*(B \cap A \cap E) + P^*(((B^c \cap A) \cup (B \cap A^c)) \cup (B^c \cap A^c)) \cap E)
\]

Now notice that

\[
(B^c \cap A) \cup (B \cap A^c) \cup (B^c \cap A^c) = ((B^c \cap A) \cup (B^c \cap A^c)) \cup ((B^c \cap A^c) \cup (B \cap A^c))
\]

\[
= B^c \cup A^c = (B \cap A)^c
\]

$\square$
Lemma 2.7. If the sets $A_n \in \mathcal{M}$ are disjoint then

\begin{equation}
P^*(E \cap \bigcup_n A_n) = \sum_n P^*(E \cap A_n)
\end{equation}

Note that we do not yet know whether $\bigcup A_n \in \mathcal{M}$, but the formula makes sense as $P^*$ is a function on $2^\Omega$.

**Proof.** Consider first the case of a finite number of sets $A_1, \ldots, A_n$. WLOG, $n \geq 2$. Given disjoint $A_1, A_2$, write $E \cap (A_1 \cup A_2) = (E \cap (A_1 \cup A_2) \cap A_1) \cup (E \cap (A_1 \cup A_2) \cap A_2^c)$ and use definition (2.2) with $E$ replaced by $E \cap (A_1 \cup A_2)$. This gives

\begin{equation}
P^*(E \cap (A_1 \cup A_2)) = P^*(E \cap (A_1 \cup A_2) \cap A_1) + P^*(E \cap (A_1 \cup A_2) \cap A_2^c)
\end{equation}

Noting that $A_1, A_2$ are disjoint, we have $E \cap (A_1 \cup A_2) \cap A_1 = E \cap A_1$ and $E \cap (A_1 \cup A_2) \cap A_2^c = E \cap A_2$, so (2.4) hold for $n = 2$ sets.

Since $\mathcal{M}$ is a field, induction now shows that (2.4) hold for $n$ sets: $P^*(E \cap \bigcup_{k=1}^n A_k) = P^*(E \cap \bigcup_{k=1}^{n-1} A_k)$

Now we use monotonicity:

\begin{equation}
P^*(A \cap \bigcup_{k=1}^\infty A_k) \geq P^*(A \cap \bigcup_{k=1}^n A_k) = \sum_{k=1}^n P^*(E \cap E_k)
\end{equation}

and we let $n \to \infty$. The reverse inequality follows by subadditivity Proposition 2.5. \hfill \Box

Lemma 2.8. $\mathcal{M}$ is a $\sigma$-field. Set function $P : \mathcal{M} \to \mathbb{R}$ defined by $P(A) = P^*(A)$ is a probability measure.

Proof. By (2.4) used with $E = \Omega$, $P^*$ restricted to $\mathcal{M}$ is countably additive. However, we do not apriori know whether $\bigcup_n A_n \in \mathcal{M}$.

Suppose $A_1, A_2, \ldots$ are disjoint with $A = \bigcup_n A_n$. Then $F_n = \bigcup_{k=1}^n A_n \in \mathcal{M}$ (field), so $P^*(E) = P^*(E \cap F_n) + P^*(E \cap F_n^c)$. Applying (2.4) to the first term and monotonicity to the second term we get $P^*(E) \geq \sum_{k=1}^n P^*(E \cap A_k) + P^*(E \cap A^c)$.

Now let $n \to \infty$ and use (2.4) to see that $P^*(E) \geq P^*(E \cap A) + P^*(E \cap A^c)$. Using again subadditivity, this shows that $A \in \mathcal{M}$.

Thus $\mathcal{M}$ is closed under the countable unions of disjoint sets. It remains to prove the following lemma. \hfill \Box
Lemma 2.9. If $M$ is a field and is closed under countable unions of disjoint sets then it is a $\sigma$-field.

Proof. Given a collection of sets $\{A_n\}$ in $M$ construct sets $B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1})$. It is clear that $B_n \in M$ are disjoint and $\bigcup_n A_n = \bigcup_n B_n$.

To conclude the proof, we need to show that $F_0 \subseteq M$ so that $F = \sigma(F_0) \subseteq M$.

Lemma 2.10. $F_0 \subseteq M$

Proof. Let $A \in F_0$. In view of subadditivity, we only need to verify that (2.3) holds for every $E \subseteq \Omega$.

Fix $\varepsilon > 0$ and let $A_n \in F_0$ be such that $E \subseteq \bigcup A_n$ and $\varepsilon + P^*(E) > \sum_n P(A_n)$.

Since $A \cap E \subseteq \bigcup_j (A \cap A_n)$ and $A^c \cap E \subseteq \bigcup_j (A^c \cap A_n)$, we have $P^*(A \cap E) + P^*(A^c \cap E) \leq \sum_n P(A \cap A_n) + \sum_n P(A^c \cap A_n)$. By finite additivity, $P^*(A \cap E) + P^*(A^c \cap E) \leq \sum_n P(A_n) < P(E) + \varepsilon$.

We can now complete the proof of Theorem. Since $P$ and $P^*$ coincide on $M$ and $P^*$ and $P_0$ coincide on $F_0$, we already know that $P$ and $P_0$ coincide on $F_0$. Since $F_0 \subseteq M$, therefore it is also countably additive on a smaller $\sigma$-field $F$ generated by the field $F_0$.

Remark 2.11. $P_*(A) = 1 - P^*(A)$ is called the inner measure. [Billingsley] gives other expressions for the outer and inner measures which are of importance in the theory of stochastic processes.

[Do we want anything about approximations?]

Remark 2.12. For every $A \in F$ and every $\varepsilon > 0$, there exists $B \in F_0$ such that $P((A \setminus B) \cup (B \setminus \ A)) < \varepsilon$.

Proof. Fix $A \in F$. We use here that by the proof of Caratheodory’s theorem, $P(A) = P^*(A)$. In view of (2.1), for every $\varepsilon > 0$ there exists a countable collection of disjoint sets $B_j \in F_0$ such that $A \subseteq \bigcup_{n=1}^\infty B_n$ and $P(A) \leq P(\bigcup_{n=1}^\infty B_n) < P(A) + \varepsilon/2$. And then there exists $n$ such that $P(\bigcup_{k=1}^n B_k) < P(\bigcup_{n=1}^\infty B_n) + \varepsilon/2$. So with $B = \bigcup_{k=1}^\infty B_k$ we have

$$P((A \setminus B) \cup (B \setminus A)) \leq P(A \setminus B) + P(B \setminus A) \leq P(\bigcup_{n=1}^\infty B_n \setminus B) + P(\bigcup_{n=1}^\infty B_n \setminus A) < \varepsilon/2 + \varepsilon/2$$

□

2. Uniqueness

This section is based on [Billingsley, Section 3].

Theorem 2.13. A (countably additive) probability measure on a field has a unique extension to the generated $\sigma$-field.

In view of Theorem 2.3, we only need to prove uniqueness. This is accomplished using some more theory, which extracts appropriate property of the field, and combines it with “natural property” of the sets that two measures coincide. This theory yields the proof on page 16.
2.1. Dynkin’s \( \pi \)-\( \lambda \) Theorem.

**Definition 2.1.** A class \( \mathcal{P} \) of subsets of \( \Omega \) is a \( \pi \)-system if

\[
(\pi) \text{ } A, B \in \mathcal{P} \text{ implies } A \cap B \in \mathcal{P}.
\]

Examples of \( \pi \)-systems are

(i) \( \{\emptyset\} \), which generates sigma-field ...

(ii) Family of intervals \(({-\infty}, a]\) with \(a \in \mathbb{R}\), which generates Borel sigma-field \(B_{\mathbb{R}}\)

(iii) Family \(({-\infty}, a] \times (-\infty, b]\), which generates Borel sigma field \(B_{\mathbb{R}^2}\)

(iv) Family of sets \(B_1 \times B_2 \times \cdots \times B_d \times \mathbb{R}_\infty\) with \(B_j \in B_{\mathbb{R}}\) which generates the Borel sigma field \(B_{\mathbb{R}_\infty}\).

**Definition 2.2.** A class \( \mathcal{L} \) of subsets of \( \Omega \) is a \( \lambda \)-system if

\[
(\lambda_1) \text{ } \Omega \in \mathcal{L}.
\]

\[
(\lambda_2) \text{ } A \in \mathcal{L} \text{ implies } A^c \in \mathcal{L}.
\]

\[
(\lambda_3) \text{ If } A_1, A_2, \ldots, A_n, \ldots \in \mathcal{L} \text{ are (pairwise) disjoint then } \bigcup_n A_n \in \mathcal{L}.
\]

**Remark 2.14.** From \((\lambda_1)\) and \((\lambda_2)\) we see that \(\emptyset \in \mathcal{L}\). So if \(A, B \in \mathcal{L}\) are disjoint then \((\lambda_3)\) we get \(A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \cdots \in \mathcal{L}\).

Of course, every field is a \( \pi \)-system, and every \( \sigma \)-field is a \( \lambda \)-system.

**Lemma 2.15.** A class of sets that is both a \( \pi \)-system and a \( \lambda \)-system is a \( \sigma \)-field.

**Proof.** Clearly, if \( \mathcal{F} \) is a \( \lambda \)-system and a \( \pi \) system then it is a field. Suppose \( A_n \in \mathcal{F} \). Then \( B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1}) = A_n \cap A_1^c \cap \cdots \cap A_{n-1}^c \in \mathcal{F} \), too. We note that \( \bigcup_n A_n = \bigcup_n B_n \in \mathcal{F} \) as a disjoint sum. \( \square \)

(Omitted in 2019)

**Lemma 2.16.** Suppose \( \mathcal{P} \) is a \( \pi \)-system and \( \mathcal{L}_0 \) is the \( \lambda \)-system generated by \( \mathcal{P} \). Then \( \mathcal{L}_0 \) is a \( \sigma \)-field.

**Sketch of proof.** Because of Lemma 2.15, to show that \( \mathcal{L}_0 \) is a \( \sigma \)-field it is enough to show that it is a \( \pi \)-system. That is, we need to show that \( A, B \in \mathcal{L}_0 \) implies \( A \cap B \in \mathcal{L}_0 \).

This is done in two steps: first fix \( A \in \mathcal{P} \) and look at the collection \( \mathcal{C}_A \) of all sets \( B \) such that \( A \cap B \in \mathcal{L}_0 \). This collection turns out to be a \( \lambda \)-system\(^1\). Since \( \mathcal{P} \subset \mathcal{C}_A \), we have \( \mathcal{L}_0 \subset \mathcal{C}_A \). And this holds for any \( A \in \mathcal{P} \). This shows that if \( A \in \mathcal{P} \) and \( B \in \mathcal{L}_0 \) then \( A \cap B \in \mathcal{L}_0 \).

Now fix \( B \in \mathcal{L}_0 \) and look at the collection \( \mathcal{C}_B \) of all sets \( A \) such that \( A \cap B \in \mathcal{L}_0 \). By the previous part, \( \mathcal{P} \subset \mathcal{C}_B \). Again, \( \mathcal{C}_B \) turns out to be a \( \lambda \)-system, so \( \mathcal{L}_0 \subset \mathcal{C}_B \). This proves the lemma: for every \( B \in \mathcal{L}_0 \) and every \( A \in \mathcal{L}_0 \) we have \( A \cap B \in \mathcal{L}_0 \).

It remains to prove that the collections of sets \( \mathcal{C}_A \) and \( \mathcal{C}_B \) are \( \lambda \)-systems. This proof is omitted. \( \square \)

**Theorem 2.17** (Dynkin’s \( \pi \)-\( \lambda \) Theorem). *Suppose a \( \lambda \)-system \( \mathcal{L} \) includes a \( \pi \)-system \( \mathcal{P} \). Then \( \sigma(\mathcal{P}) \subset \mathcal{L} \).*
Proof. Let $\mathcal{L}_0$ be a $\lambda$-system generated by $\mathcal{P}$. Then $\mathcal{P} \subset \mathcal{L}_0 \subset \mathcal{L}$. From Lemma 2.16 we know that $\mathcal{L}_0$ is a $\sigma$-field and it contains $\mathcal{P}$. So $\sigma(\mathcal{P}) \subset \mathcal{L}_0 \subset \mathcal{L}$. □

Proposition 2.18. Let $\mathcal{P}$ be a $\pi$-system and denote $\mathcal{F} = \sigma(\mathcal{P})$. Suppose $P_1, P_2$ are two probability measures on $\mathcal{F}$ that agree on $\mathcal{P}$. Then $P_1 = P_2$ (on $\mathcal{F}$.)

Proof. Let $\mathcal{L}$ be the family of all sets in $\mathcal{F}$ on which $P_1$ and $P_2$ agree. Then $\mathcal{L}$ is a $\lambda$-system. By Theorem 2.17 $\mathcal{F} \subset \mathcal{L}$. □

Proof of Theorem 2.13. A field $\mathcal{F}_0$ is a $\pi$-system. So if $P_1(A) = P_2(A)$ for all $A \in \mathcal{F}_0$, then by Proposition 2.18 the same holds for all $A \in \mathcal{F} = \sigma(\mathcal{F}_0)$. □

3. Probability measures on $\mathbb{R}$

This is based on [Billingsley, Section 12] and [Durrett, Section 1.2].

Definition 2.3. $F : \mathbb{R} \rightarrow \mathbb{R}$ is a cumulative distribution function, if

(i) $F$ is non-decreasing: $x < y$ implies $F(x) \leq F(y)$
(ii) $\lim_{x \to \infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1.$
(iii) $F$ is right-continuous, $\lim_{x \to x_0^+} F(x) = F(x_0)$

Suppose that $P$ is a probability measure on the Borel subsets of $\mathbb{R}$. Consider a function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = P((−\infty, x])$. Then $F$ is a cumulative distribution function. (You should be able to supply the proof!)

The following is a combination of Lebesgue’s Theorem 1.6, with Caratheodory’s Theorem 2.3 and uniqueness Theorem 2.13.

Proposition 2.19. Every cumulative distribution function $F$ corresponds to a unique probability measure $P$ on the Borel sigma-field set of $\mathbb{R}$, such that $F(x) = P((−\infty, x])$.

Proof. Intervals of the form $(−\infty, a]$ form a $\pi$-system, and generate the Borel $\sigma$-field. So uniqueness follows from Theorem 2.13.

Consider the field $\mathcal{B}_0$ of finite disjoint unions of intervals $(a, b]$ where $−\infty \leq a < b \leq \infty$.

For finite $a < b$, define $P((a, b]) = F(b) − F(a)$. Also define $P((−\infty, a]) = F(a)$ and $P((a, \infty)) = 1 − F(a)$.

Extend $P$ by additivity to $\mathcal{B}_0$. As in Theorem 1.5, one needs to show that this definition is consistent, that $P$ is finitely-additive, and that $P$ is continuous (countably-additive) on $\mathcal{B}_0$. Once we prove this, we invoke Theorem 2.3.
3. Probability measures on \( \mathbb{R} \)

(3.1.1. Uniform distributions.)

3.1.1. Uniform distributions.

Example 2.1 (Uniform I). Uniform distribution on the set of real numbers \( \{x_1 < x_2 < \cdots < x_n\} \) is (see Examples 1.3 and 1.4) \( P = \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j} \) and corresponds to \( F(x) = \# \{ j : x_j \leq x \} / n \).

Example 2.2 (Uniform II). Uniform distribution on the interval \((0,1)\) is the probability measure \( P \) which corresponds to

\[
F(x) = \begin{cases} 
0 & x < 0 \\
 x & \text{for } 0 \leq x \leq 1 \\
1 & x > 1
\end{cases}
\]

Notation: \( U(0,1) \). More generally, \( U(a,b) \) corresponds to \( F(x) = (x-a)/(b-a)1_{(a,b]} + 1_{[b,\infty)} \).

Recall the construction of the Cantor set: split \([0,1]\) into \([0,1/3) \cup (1/3,2/3) \cup [2/3,1]\) and remove the middle part. Continue recursively the same procedure with each of the closed intervals retained.
Example 2.3 (Uniform III). Uniform distribution on the Cantor set corresponds to $F$ that is constant on all deleted intervals,

$$F(x) = \begin{cases} 
0 & x < 0 \\
\vdots \\
1/4 & 1/9 \leq x < 2/9 \\
\vdots \\
1/2 & 1/3 \leq x < 2/3 \\
\vdots \\
3/4 & 7/9 \leq x < 8/9 \\
\vdots \\
1 & x \leq 1 
\end{cases}$$

The interval removed in $d$-th step is $(\sum_{k=1}^{d} x_k/3^k, \sum_{k=1}^{d} x_k/3^k+1/3^d)$ with $x_d = 1$ and $x_1, \ldots, x_{k-1} \in \{0, 2\}$. For example, for $d = 1$ it is $(1/3, 1/3 + 1/3)$. For $d = 2$ the intervals are $(1/3^2, 1/3^2 + 1/3^2)$ and $(2/3 + 1/3^2, 2/3 + 1/3^2 + 1/3^2)$. On each removed interval, $F(x) = \sum_{k=1}^{d-1} x_k/2^{k+1} + 1/2^d$ is constant.

![Graph of F(x)](image)

3.1.2. Important (absolutely) continuous distributions. Continuous distributions arise from $F(x) = \int_{-\infty}^{x} f(y)dy$, where the so called density function $f \geq 0$ and $\int_{-\infty}^{\infty} f(y)dy = 1$. Example 2.2 is absolutely continuous with $f(y) = 1_{[a,b]}$.

Example 2.4 (Exponential distribution). Take $f(x) = \lambda e^{-\lambda x}I_{(0,\infty)}(x)$, where $\lambda > 0$. This gives

$$F(x) = \begin{cases} 
0 & x < 0 \\
1 - e^{-\lambda x} & x \geq 0.
\end{cases}$$

Example 2.5 (Standard normal distribution). Take $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Notation: $N(0,1)$. 


3.1.3. **Other examples.**

**Example 2.6** (mixed type). It is clear that

\[
F(x) = \begin{cases} 
0 & x < 0 \\
\frac{x}{9} & 0 \leq x < 1 \\
\frac{x}{3} & 1 \leq x < 2 \\
1 & x \geq 2 
\end{cases}
\]

is a cumulative distribution function which cannot be written as an integral of a density.

4. **Probability measures on \( \mathbb{R}^k \)**

For simplicity consider only \( k = 2, 3 \).

4.1. **Probability measures on \( \mathbb{R}^2 \).** The \( \pi \) system that generates Borel sets of \( \mathbb{R}^2 \) consists of sets \( (-\infty, x] \times (-\infty, y] \). Thus every probability measure \( P \) on Borel sets of \( \mathbb{R}^2 \) is determined uniquely by its values on such sets, \( F(x, y) = P((-\infty, x] \times (-\infty, y]) \). Function \( F(x, y) \) is called a joint cumulative distribution function.

The probability measure must assign nonnegative numbers to all rectangles \( A = (a_1, b_1] \times (a_2, b_2] \). It is clear (draw a picture) that

\[
(-\infty, b_1] \times (-\infty, b_2] = (-\infty, a_1] \times (-\infty, b_2] \cup (-\infty, b_1] \times (-\infty, a_2] \cup A
\]

Thus

\[
F(b_1, b_2) = P(A) + P((-\infty, a_1] \times (-\infty, b_2] \cup (-\infty, b_1]) \\
= P(A) + F(a_1, b_2) + F(a_2, b_1) - P((-\infty, a_1] \times (-\infty, b_2) \cap (-\infty, b_1]) \\
= P(A) + F(a_1, b_2) + F(a_2, b_1) - F(a_1, a_2)
\]

Thus

\[
P(A) = \Delta_A(F) := F(b_1, b_2) + F(a_1, a_2) - F(a_1, b_2) - F(a_2, b_1)
\]

\footnote{Probability measures of mixed type arise in actuarial models, where the loss of an insured person might have a density but the insurance payoff may be capped, or be a fraction of the of loss that changes when the loss exceeds some predefined thresholds.}
This shows that we must have \( \Delta_A F \geq 0 \).

It is also clear that we have the following properties:

- \( F \) is "right-continuous": if \( a_n, b_n > 0 \) converge to 0 then \( F(x + a_n, y + b_n) \to F(x, y) \).
- \( \lim_{x,y \to \infty} F(x, y) = 1 \)
- \( \lim_{y \to -\infty} F(x, y) = \lim_{x \to -\infty} F(x, y) = 0 \)
- \( G(x) = \lim_{y \to \infty} F(x, y) \) and \( H(y) = \lim_{x \to \infty} F(x, y) \) exist and define non-decreasing functions, called the marginal cumulative distribution functions.

This motivates the following definition:

**Definition 2.4.** \( F(x, y) \) is a bivariate cumulative distribution function, if the following conditions hold:

(i) \( \Delta_A F \geq 0 \) for all \( A = (a_1, a_2] \times (b_1, b_2] \)

(ii) \( \lim_{x,y \to \infty} F(x, y) = 1 \)

(iii) \( \lim_{y \to -\infty} F(x, y) = \lim_{x \to -\infty} F(x, y) = 0 \)

(iv) \( F \) is right-continuous,

The following is an analog of Proposition 2.19.

**Proposition 2.20.** Every cumulative distribution function \( F(x, y) \) corresponds to a unique probability measure.

**Sketch of proof.** The field \( \mathcal{B}_0 \) generated by the sets \( (-\infty, b_1] \times (-\infty, b_2] \) consists of finite unions of disjoint sets that arise as intersections of such sets or their complements, see Exercise 1.16.

This gives sets \( (-\infty, b_1] \times (-\infty, b_2] \), their complements, finite rectangles \( A \), sets of the form \( (-\infty, b_1] \times (a_2, b_2] \) and \( (a_1, b_1] \times (-\infty, b_2] \).

We define \( P((a_1, \infty) \times (a_2, \infty)) = 1 - F(a_1, a_2) \), \( P(A) = \Delta_A F \), \( P((-\infty, b_1] \times (-\infty, b_2]) = F(b_1, b_2) \) and \( P((-\infty, b_1] \times (a_2, b_2]) = \lim_{a_1 \to -\infty} \Delta_A F \). We extend the definition by additivity to \( \mathcal{B}_0 \).

Next we check that the assumptions of Exercise 1.11 are again satisfied, so we can conclude that \( P \) has a unique countably additive extension to the Borel \( \sigma \)-field.

It suffices to find a suitable compact set for each of the four types of the "generalized" rectangles. If \( A = (a_1, \infty) \times (a_2, \infty) \) we take \( K = [a_1 + \delta, B_1] \times [a_2 + \delta, B_2] \) and \( B = (a_1 + \delta, B_1] \times (a_2 + \delta, B_2] \).

Given \( \varepsilon > 0 \) choose \( \delta \) such that \( F(a_1 + \delta, a_2 + \delta) < F(a_1, a_2) + \varepsilon \) \( B_1, B_2 \) such that \( F(B_1, B_2) > 1 - \varepsilon \).

\[ P(B) = F(B_1, B_2) + F(a_1 + \delta, a_2 + \delta) - F(a_1 + \delta, B_2) - F(a_2 + \delta, B_1) \]

\[ \square \]

**Example 2.7.** Uniform distribution on the unit square is defined by

\[
F(x, y) = \begin{cases} 
xy & 0 \leq x \leq 1, 0 \leq y \leq 1 \\
x & 0 \leq x \leq y, y > 1 \\
y & x > 1, 0 \leq y \leq 1 \\
1 & x > 1, y > 1 \\
0 & \text{otherwise}
\end{cases}
\]
4.2. Probability measures on $\mathbb{R}^3$. The $\pi$ system that generates Borel sets of $\mathbb{R}^3$ consists of sets $(-\infty,x] \times (-\infty,y] \times (-\infty,z]$. Thus every probability measure is determined uniquely by its values on such sets, $F(x,y,z)$.

We need to assign values of the measure to all rectangles $A = (a_1,b_1] \times (a_2,b_2] \times (a_3,b_3]$. It is clear that

\begin{equation}
(-\infty,b_1] \times (-\infty,b_2] \times (-\infty,b_3] = A \cup (-\infty,a_1] \times (-\infty,b_2] \times (-\infty,b_3] \cup (-\infty,b_1] \times (-\infty,a_2] \times (-\infty,b_3] \cup (-\infty,b_1] \times (-\infty,b_2] \times (-\infty,a_3]
\end{equation}

Noting that $A$ is disjoint with the remaining set, by the inclusion-exclusion formula (1.2), we get

\begin{equation}
F(b_1,b_2,b_3) = P(A) + F(a_1,b_2,b_3) + F(b_1,a_2,b_3) + F(b_1,b_2,a_3) - F(a_1,a_2,b_3) - F(a_1,b_2,a_3) - F(b_1,a_2,a_3) + F(a_1,a_2,a_3)
\end{equation}

So

\begin{equation}
P(A) = \Delta_A(F) := F(b_1,b_2,b_3) + F(a_1,b_2,b_3) + F(a_1,b_2,a_3) + F(b_1,a_2,b_3) - F(a_1,a_2,b_3) - F(b_1,a_2,a_3) - F(b_1,b_2,a_3) - F(a_1,a_2,a_3)
\end{equation}

An analog of Definition 2.4 uses $\Delta_A(F)$ as defined in (2.9). Proposition 2.20 has an $\mathbb{R}^3$ version. Similar approach works in $k$ dimensions, compare [Durrett, Theorem 1.1.6] or [Billingsley, Theorem 12.5], who consider general measures. (In general, $\Delta_A(F)$ is defined using the inclusion-exclusion principle (1.2). Note that for unbounded measures $F$ can take negative values!)

4.3. Probability measures on $\mathbb{R}^\infty$. Recall that $\mathbb{R}^\infty$ is the set of all infinite real sequences, with metric (A.4). Probability measures on $\mathbb{R}^\infty$ are determined uniquely by the families of joint finite-dimensional distributions that arise from a special $\pi$-system of cylindrical sets, i.e. sets of the form

\begin{equation}
(-\infty,a_1] \times (-\infty,a_2] \times \cdots \times (-\infty,a_n] \times \mathbb{R} \times \mathbb{R} \times \ldots.
\end{equation}

A special case of such a measure is constructed in Theorem 4.10. This is one place where probability theory “outperforms” the general measure theory - while there is a Lebesgue measure on $\mathbb{R}^d$, there is no Lebesgue measure on $\mathbb{R}^\infty$.

4.4. Probability measures on $\Omega = C[0,1]$. \(^3\) Constructions of probability measures on function spaces such as $C[0,1]$ usually rely on the $\pi$ system of sets of the form $\{f : f(t_1) \leq x_2, \ldots, f(t_n) \leq x_n\}$ which are indexed by $t_1,\ldots,t_n \in [0,1]$ and $x_1,\ldots,x_n \in \mathbb{R}$. These are sometimes referred to as cylindrical sets.

The functions $F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \Pr(f(t_1) \leq x_2, \ldots, f(t_n) \leq x_n)$ are called the finite dimensional distributions. For fixed $t_1,\ldots,t_n$, $F_{t_1,\ldots,t_n}(x_1,\ldots,x_n)$ is a cumulative distribution function which determines a family of probability measures $P_{t_1,\ldots,t_n}$ on Borel subsets of $\mathbb{R}^n$. These measures determine a probability measure $\Pr$ on $C[0,1]$ uniquely, but it is easy to see that to do so they must be “consistent”. An example of a consistency condition is $P_{t_1}(A) = P_{t_1,t_2}(A \times \mathbb{R})$.

Constructions of such measures requires good understanding of compact subsets of $C[0,1]$. 

(Omitted in 2019)
4.5. Probability measures on $\Omega = \mathbb{R}^{[0,1]}$. Since compact sets are easy to find in product spaces, the simplest example of a probability measure on an infinite dimensional space is the case of $\Omega = \mathbb{R}^{[0,1]}$.

**Theorem 2.21** (Kolmogorov). Suppose probability measures $P_{t_1}, \ldots, t_n$ are consistent. Then there exists a unique probability measure $\Pr$ on $\mathbb{R}^{[0,1]}$ with Borel $\sigma$-field that generates $P_{t_1}, \ldots, t_n$ as finite dimensional distributions.

**Remark 2.22.** A good description of Borel $\sigma$-field in $\mathbb{R}^{[0,1]}$ appears in [Billingsley, Section 36]. In particular, the subset $C^{[0,1]} \subset \mathbb{R}^{[0,1]}$ is not a Borel set! However, for a given $\Pr$ one can ask what is $\Pr^*$ and $\Pr^*$ of $C^{[0,1]}$.

Good probability measures are those for which $\Pr^*(C^{[0,1]}) = 1$ and $\Pr^*(C^{[0,1]}) = 0$.

**Proof.** The steps in the proof are:
- Introduce the field $F_0$ of cylindrical sets, indexed by $t_1, \ldots, t_n$ and Borel subsets of $\mathbb{R}^n$.
- Define a probability measure $\Pr$ on $F_0$ by using the finite-dimensional distributions $P_{t_1}, \ldots, t_n$.
- One then uses a variant of the compactness argument similar to Exercise 1.11 to verify that if $A_n$ is a decreasing family of sets in $F_0$ with $\bigcap_n A_n = \emptyset$ then $\Pr(A_n) \to 0$.

$\square$

5. Lebesgue measure

Occasionally we need to consider positive measures on a $\sigma$-field that are not probability measures. The main example is the Lebesgue measure $\lambda$ on $(\mathbb{R}, \mathcal{B})$, which corresponds to strictly increasing and continuous but unbounded function $F(x) = x$ and the Lebesgue measure $\lambda$ on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$.

6. Transport of measure

Suppose we have two measure spaces $\Omega_1, F_1$ and $\Omega_2, F_2$.

**Definition 2.5.** We say that a function $\varphi : \Omega_1 \to \Omega_2$ is measurable if for every $B \in F_2$ its inverse image $\varphi^{-1}(B) \in F_2$.

Recall that $\varphi^{-1}(B) = \{\omega \in \Omega_1 : \varphi(\omega) \in B\}$.

A transport of measure is a construction of a probability measure on $(\Omega_2, F_2)$ from measure $P_1$ and function $\varphi$. This topic will return in Definition 4.3, where $P_2$ is called the induced measure. The definition of $P_2$ is as follows. For $B \in F_2$, we define

$$P_2(B) := P_1(\varphi^{-1}(B))$$

It is known that a continuous function $(0,1) \to \mathbb{R}$ is measurable. We will show this in Proposition 4.5 on page 38, and for now we use it without proof to give examples with $\Omega_1 = (0,1)$ with Lebesgue measure $\lambda$.

**Example 2.8.** If $f : (0,1) \to \mathbb{R}$ is $\varphi(t) = \ln t$ then the CDF of $P_2$ is

$$P_2((-\infty, x]) = \lambda(\{t \in (01) : \varphi(t) \leq x\}) = \lambda(\{t \in (01) : \leq e^x\}) = \lambda((0,1) \cap (-\infty, e^x]) = \begin{cases} 1 & x \geq 0 \\ e^x & x < 0 \end{cases}$$

Measure $P_2$ has density

$$f(x) = \begin{cases} 0 & x > 0 \\ e^x & x < 0 \end{cases}$$
Example 2.9. If \( \varphi : (0, 1) \to \mathbb{R} \) is \( \varphi(t) = \tan(\pi t - \pi/2) \) then a similar calculation shows that the CDF of the induced measure \( P_2 \) on \( (\mathbb{R}, \mathcal{B}) \) is \( F_2(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x \). The corresponding density
\[
f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}
\]
is known as the Cauchy density.

Required Exercises

Exercise 2.1 (Different representations of the same probability measure on a \( \sigma \)-field \( \mathcal{F} \)). Let \( \lambda \) be the Lebesgue measure on the Borel \( \sigma \)-field of subsets of \( \Omega = [0, 1] \). Consider \( \pi \)-system \( \mathcal{P} = \{ [0, 1/n] : n \in \mathbb{N} \} \) and let \( \mathcal{F} = \sigma(\mathcal{P}) \). Show that there exists a discrete probability measure \( P = \sum_{n=1}^{\infty} p_n \delta_{\omega_n} \) on \( 2^{[0,1]} \) (see Example 1.4) such that \( \lambda \) restricted to \( \mathcal{F} \) coincides with \( P \) restricted to \( \mathcal{F} \). (In formal notation, \( \lambda|_{\mathcal{F}} = P|_{\mathcal{F}} \).)

Solution: Measure \( P = \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \delta_{1/n} \) is a discrete probability measure with the property that \( P((0, 1/n)) = \sum_{k=n+1}^{\infty} \frac{1}{k(k-1)} = \sum_{k=n+1}^{\infty} (\frac{1}{k-1} - \frac{1}{k}) = 1/n \). Since both \( P \) and \( \lambda \) agree on a \( \pi \)-system, they must agree on the sigma field generated by this \( \pi \)-system. So indeed \( \lambda|_{\mathcal{F}} = P|_{\mathcal{F}} \). One can set up similar discrete measures, so this \( P \) is not unique.

Exercise 2.2 (Statistics). It is illustrative to produce empirical histograms at various sample sizes for the uniform distribution on the Cantor set from Example 2.3. Somewhat surprisingly, this distribution is easy to simulate by taking \( 2^{\sum_{k=1}^{\infty} \varepsilon_k / 3^k} \) where \( \varepsilon_k \) represents a “toss of a fair coin” with values 0 or 1. This exercise asks you to reproduce some of the histograms that appear in [Proshan-Shaw].

Solution: This is a computer exercise!

Exercise 2.3 (measure-preserving maps). Let \( \varphi : [0, 1] \to [0, 1] \) be the fractional part of \( 2x \). That is,
\[
\varphi(x) = \begin{cases} 
2x & \text{if } x \leq 1/2 \\
2x - 1 & \text{if } x > 1/2
\end{cases}
\]
Show that for every Borel subset \( A \) of \( [0, 1] \) the Lebesgue measure of \( \varphi^{-1}(A) \) equals to the Lebesgue measure of \( A \). (Compare Exercise 2.12.)

Solution: Consider a \( \pi \)-system of diadic intervals \( I = [0, \frac{k}{2^n}] \) with \( k, n \in \mathbb{N} \). Then \( f^{-1}(I) = \{ x \leq 1/2 : 2x \in [0, \frac{k}{2^n}] \} \cup \{ x > 1/2 : 2x - 1 \in [0, \frac{k}{2^n}] \} = [0, k/2^{n+1}] \cup (1/2, 1/2 + k/2^{n+1}] \) is the union of two disjoint intervals of length \( k/2^{n+1} \). Therefore, the length of \( I \) is preserved. Now introduce an appropriate \( \lambda \)-system, and use the \( \pi - \lambda \) theorem.

Exercise 2.4. What should be the CDF \( F(x, y) \) for the distribution “uniform on the triangle” \( x \geq 0, y \geq 0, x + y \leq 1? \)
Solution: It is easy to overlook some of the six cases, so it's best to make a picture.

Here are the six cases:

\[
F(x,y) = \begin{cases} 
1 & x \geq 1, y \geq 1 \\
1 - 2(1-y)^2 & x > 1, 0 < y < 1 \\
? & 0 < x < 1, y > 1 \\
2xy & 0 < x < 1, 0 < y < 1, x + y < 1 \\
? & 0 < x < 1, 0 < y < 1, x + y > 1 \\
0 & x \leq 0 \text{ or } y \leq 0
\end{cases}
\]

Exercise 2.5. Consider \((\Omega_1, \mathcal{F}, \mathbb{P}_1) = ((0,1), \mathcal{B}, \lambda)\) and measurable function \(\varphi : (0,1) \to \mathbb{R}\) given by \(\varphi(t) = 4t(1-t)\). Find the CDF for the measure induced by \(\varphi\) on \((\mathbb{R}, \mathcal{B})\).

Additional Exercises

Exercise 2.6. Let \(\Omega = (0,1] \times (0,1]\) and let \(\mathcal{F}\) be the class of sets of the form \(A_1 \times (0,1]\) with \(A_1 \in \mathcal{B}\) the Borel \(\sigma\)-field in \((0,1]\) and \(P(A_1 \times (0,1]) = \lambda(A_1)\) (the Lebesgue measure). Then \((\Omega, \mathcal{F}, \mathbb{P})\) is a probability space. For the diagonal \(D = \{(x,x) : 0 < x \leq 1\}\), find \(P^* (D)\) and \(P^* (D^c)\).

Solution: \(P^*(D) = \inf \sum \lambda(A_j) : \bigcup A_j \times (0,1] \supset D\). This means that \(\bigcup A_j = (0,1]\), so the smallest possible value of \(\sum \lambda(A_j)\) occurs when \(A - j\) are disjoint and cover \((0,1]\). So the answer is \(P^*(D) = 1\). Clearly \(D^c\) also cannot be covered by sets \(A_j \times (0,1]\) unless \(A_j\) alone cover \((0,1]\). So \(P^*(D^c) = 1\), too.

Exercise 2.7. Inspect the proofs of Theorems 2.3 and 2.13. Find all places where additivity or countable additivity is used.

Solution: The proofs of these theorems were omitted in 2018!

Exercise 2.8 (Compare Exercise 1.17). For \(\Omega = (0,1]\) with the field \(\mathcal{B}_0\) generated by intervals \(I = (a,b]\), consider \(\lambda_0(I) = |I|\), extended by additivity to \(\mathcal{B}_0\). Let \(Q\) be the set of all rational numbers in \((0,1]\) Use the definition of \(\lambda^*\) (not subadditivity) to show that \(\lambda^*(Q) = 0\).
Solution: By definition, \( \lambda^*(\mathbb{Q}) = \inf \sum \lambda(I_k) : \bigcup I_k \supset Q \}. \) Choose \( \varepsilon > 0 \) and let \( q_1, q_2, \ldots \) be the sequence of all rational numbers in \( Q = \mathbb{Q} \cap (0, 1] \). Take \( I_k = (q_k - \varepsilon/2^k, q_k + \varepsilon/2^k] \cap (0, 1] \). Then \( \lambda(I_k) \leq 2\varepsilon/2^k \) so \( \lambda^*(\mathbb{Q}) \leq \sum_{k=1}^{\infty} 2\varepsilon/2^k = 2\varepsilon \). Since \( \varepsilon > 0 \) was arbitrary, this shows that \( \lambda^*(\mathbb{Q}) = 0 \).

Exercise 2.9. The family \( \mathcal{P} \) of open intervals \((−1/n, 1/n)\) with \( n \in \mathbb{N} \) is a \( \pi \)-system in \( \Omega = (−1, 1) \). Describe what sets are in the \( \sigma \)-field \( \sigma(\mathcal{P}) \). In particular, is set \{0\} in \( \sigma(\mathcal{P}) \)?

Solution: Trivially, \( \emptyset \) and \( (−1, 1) \) are also in \( \mathcal{F} \). Also \( \{0\} = \bigcap_{n}(−1/n, 1/n) \) must be in \( \mathcal{F} \). So sets \((−1/n, 0) \cup (0, 1/n)\) must be in \( \mathcal{F} \), too. The only other sets in \( \mathcal{F} \) are the finite unions of symmetric intervals \((−1/n, 1/m) \cup [1/m, 1/n)\) with \( m, n \in \mathbb{N} \). (The only nontrivial countable union of such sets is \( \bigcup_{m=1}^{\infty}(−1/n, 1/m) \cup [1/m, 1/n) = (−1/n, 1/n) \setminus \{0\}.\)

Exercise 2.10. Let \( \mathcal{A} \) be the smallest field generated by a \( \pi \)-system \( \mathcal{P} \) (see Exercise 1.16). Use the inclusion-exclusion formula from Exercise 1.1 to show that finitely additive probability measures that agree on \( \mathcal{P} \) must also agree on \( \mathcal{A} \).

Exercise 2.11. Suppose \( \mathcal{L} \) is a \( \lambda \)-system. Show that \( A, B \in \mathcal{L} \) and \( A \subset B \) implies that \( B \setminus A \in \mathcal{L} \).

\textit{Hint:} Show that \( (B \setminus A)^c \in \mathcal{L} \).

Exercise 2.12. Consider \( \Omega = (0, 1) \) with Lebesgue measure. Use the Dynkin’s \( \pi \)-\( \lambda \) Theorem to prove that for all Borel sub-sets \( B \) of \((0, 1/2)\) and all \( x \in (0, 1/2) \), the Lebesgue measure of \( B + x \) is the same as the Lebesgue measure of \( B \).

Solution: Fix \( 0 < x < 1/2 \). Consider the \( \pi \)-system \( \mathcal{P} \) of sets \((0, a)\) with \( a < 1/2 \). Clearly \( \lambda(x + (0, a)) = \lambda(x, a + x) = a = \lambda(0, a) \).

Next, introduce the set \( \mathcal{L} \) of Borel subsets \( B \) of \((0, 1/2)\) with the property that \( \lambda(B + x) = \lambda(B) \). This collection is a \( \lambda \)-system: if \( B \in \mathcal{L} \) then \((0, 1/2) \setminus B \in \mathcal{L} \) because \( \lambda((0, 1/2) \setminus B) = 1/2 - \lambda(B) \) and \( \lambda(x + ((0, 1/2) \setminus B)) = \lambda((x, 1/2 + x) \setminus (x + B)) = 1/2 - \lambda(x + B) = 1/2 - \lambda(B) \). Similarly, for disjoint sets \( B_n \) sets \( x + B_n \) are disjoint, so \( \lambda(x + \bigcup B_n) = \lambda(\bigcup_{n}(x + B_n)) = \sum_{n} \lambda(x + B_n) = \sum_{n} \lambda(B_n) = \lambda(\bigcup B_n) \).

Since \( \mathcal{P} \subset \mathcal{L} \), therefore Dynkin’s \( \pi \)-\( \lambda \) Theorem we have \( B_{(0,1/2)} = \sigma(\mathcal{P}) \subset \mathcal{L} \).
1. Independent events and sigma-fields

This section follows [Billingsley, Section 4].

**Definition 3.1** (Independence of events).

(i) Events $A, B$ are independent if $P(A \cap B) = P(A)P(B)$.

(ii) More generally, events $A_1, \ldots, A_n$ are independent, if for every $r \leq n$ and every choice of distinct $k_1, \ldots, k_r$

\[
P(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_r}) = P(A_{k_1})P(A_{k_2}) \cdots P(A_{k_r})
\]

(iii) An infinite sequence of events $A_1, A_2, \ldots$ is independent if the events $A_1, \ldots, A_n$ are independent for every $n$.

**Remark 3.1.** Events $A_1, \ldots, A_n$ are independent iff

\[
P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2) \cdots P(B_n)
\]

for all choices of $B_j = A_j$ or $B_j = \Omega$.

This suggests that it is in fact easier to work with the $\sigma$-fields. Let’s re-write everything from this more abstract point of view.

**Definition 3.2.**

(i) An infinite family $\{F_t : t \in T\}$ of $\sigma$-fields is independent, if every finite sub-family is independent.

(ii) A finite family $\{F_1, \ldots, F_n\}$ is independent, if for every choice of events $B_j \in F_j$ identity (3.2) holds.

(iii) Events $A_1, \ldots, A_n, \ldots$ are independent, if the generated $\sigma$-fields $F_j = \sigma(A_j)$ are independent.

(iv) Classes of sets $A_1, A_2, \ldots, A_n$ are independent, if for each choice of events $A_j$ from $A_j$ the events $A_1, \ldots, A_n$ are independent.
Example 3.1. Consider $\Omega = [0, 1]^3$ with Lebesgue measure. Then events $A = \{(x, y, z) \in \Omega : x < 1/2\}$, $B = \{(x, y, z) \in \Omega : y < 1/2\}$, $C = \{(x, y, z) \in \Omega : z < 1/2\}$ are independent.

To verify independence of $\sigma$-fields we often rely on the following.

**Theorem 3.2.** If $A_1, \ldots, A_n$ are independent $\pi$-systems then $\sigma(A_1), \ldots, \sigma(A_n)$ are independent.

(Okated in 2019)

**Proof.** Without loss of generality we may assume $\Omega \in A_j$ so the definition to use is (3.2).

Fix $B_2, \ldots, B_n$ and consider

\[ \mathcal{L} = \{B_1 \in \sigma(A_1) \text{ such that (3.2) holds}\} \]

It is easy to see that $\mathcal{L}$ is a $\lambda$-system and $A_1$ is a $\pi$-system contained in $\mathcal{L}$. So by Theorem 2.17 we see that $\sigma(A_1) \subset \mathcal{L}$.

This proves that $\sigma(A_1), A_2, \ldots, A_n$ are independent $\pi$-systems.

We now repeat the same argument for $A_2$, then $A_3$, etc. $\square$

**Corollary 3.3.** If $A_1, A_2, \ldots$ is an infinite set of independent $\pi$-systems, then $\sigma(A_1), \sigma(A_2), \ldots$ are independent.

We will need the following auxiliary result

**Lemma 3.4.** If events $A_1, A_2, \ldots$ and $C_1, C_2, \ldots$ are (jointly\(^1\)) independent then the generated $\sigma$-fields are also independent.

**Proof.** For fixed $n$ and fixed $A \in \sigma(A_1, \ldots, A_n)$ of the form $B_1 \cap B_2 \cap \cdots \cap B_n$ where $B_k$ is either $A_k$ or $\Omega$, consider $\mathcal{L} = \{C \in \mathcal{F} : P(C \cap A) = P(C)P(A)\}$. It is clear that $\mathcal{L}$ is a $\lambda$-system, see Definition 2.2.

Next, consider the family $\mathcal{P}$ of sets of the form $D_1 \cap D_2 \cap \cdots \cap D_m$ where $D_j$ is either $C_j$ or $\Omega$ and $m = 1, 2, \ldots$. Then it is clear that $\mathcal{P}$ is a $\pi$-system, see Definition 2.1.

We now show that $\mathcal{P} \subset \mathcal{L}$. If $C \in \mathcal{P}$ then by independence (used three times!) we have

\[ P(A \cap C) = P(B_1 \cap B_2 \cap \cdots \cap B_n \cap D_1 \cap D_2 \cap \cdots \cap D_m) \]
\[ = P(B_1)P(B_2)\ldots P(B_n)P(D_1)P(D_2)\ldots P(D_m) = P(A)P(C) \]

By Dynkin’s $\pi - \lambda$ theorem (Theorem 2.17) we see that $\sigma(\mathcal{P}) \subset \mathcal{L}$. Thus for any $C \in \sigma(C_1, C_2, \ldots)$ the events $C, A_1, A_2, \ldots$ are independent.

Repeating the same argument again for a fixed $C \in \sigma(C_1, C_2, \ldots)$, we see that Dynkin’s $\pi - \lambda$ theorem implies that for every $B \in \sigma(A_1, A_2, \ldots)$ events $B, C$ are independent. Thus $\sigma(A_1, A_2, \ldots)$ and $\sigma(A_1, A_2, \ldots)$ are independent. $\square$

**Example 3.2.** Suppose events $A_1, A_2, A_3, A_4$ are independent. Then the events $A = A_1 \cup A_2$ and $B = A_3 \cup A_4$ are also independent. This can be verified by elementary calculation, but we deduce this from Corollary 3.5 or from Lemma 3.4:

Consider $\mathcal{A} = \sigma(A_1, A_2)$ and $\mathcal{B} = \sigma(A_3, A_4)$. By Corollary 3.5 these $\sigma$-fields are independent. Clearly, $A \in \sigma(\mathcal{A})$ and $B \in \sigma(\mathcal{B})$.

---

\(^1\)i.e., the events $A_1, B_1, A_2, B_2, \ldots$ are independent
Later on, we will also need to consider more than two sequences of events. The following generalizes Lemma 3.4 to a setting with infinite number of sequences.

**Proposition 3.5.** If \( A_{i,j} \) is an (possibly infinite) array of independent events then the \( \sigma \)-fields generated by each row of the array are independent.

(Omitted in 2019)

**Proof.** We introduce \( \pi \)-systems \( A_i \) as the collection of all finite intersections of the events in the \( i \)-th row, including \( \Omega \). If \( C_i \in A_i \) then \( C_i = \bigcap_{k=1}^{m_i} A_{ik} \) where \( A_{ik} \in A_i \)

\[
P(\bigcap_{i=1}^{n} C_i) = P(\bigcap_{i=1}^{n} \bigcap_{k=1}^{m_i} A_{ik}) = \prod_{i=1}^{n} \prod_{k=1}^{m_i} P(A_{ik}) = \prod_{i=1}^{n} P(C_i)
\]

We now apply Corollary 3.3. \( \square \)

Another natural application of Corollary 3.3 says that in Example 3.1 independence extends to the three sigma fields generated by the "coordinates".

2. Zero-one law

**Definition 3.3.** The tail \( \sigma \)-field for a sequence of events \( A_1, A_2, \ldots \) is

\[
\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots)
\]

**Example 3.3.** \( \{ A_n \ i.o. \} = \bigcap_{n} \bigcup_{k \geq n} A_k \) and \( \{ A_n \ eventually \} = \bigcup_{n} \bigcap_{k \geq n} A_k \) are tail events.

**Proof.** \( \bigcap_{n} \bigcup_{k \geq n} A_k = \{ A_n \ i.o. \} \) is a tail event: to determine if the infinite number of events occurred, we do not need to know anything about \( A_1 \), or \( A_2 \), etc. Formally, for every \( N > 1 \) we have

\[
\bigcup_{n=1}^{\infty} \bigcup_{k \geq n} A_k = \bigcap_{n=N}^{\infty} \bigcup_{k \geq n} A_k \in \sigma(A_N, A_N + 1, \ldots)
\]

because for decreasing events \( B_n := \bigcup_{k \geq n} A_k \) we have \( B_1 \cap B_2 \cap \cdots \cap B_N = B_N \) and

\[
\bigcap_{n=1}^{\infty} B_k = B_1 \cap B_2 \cap \cdots \cap B_N \cap \bigcap_{n=N}^{\infty} B_k
\]

Thus

\[
\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k \in \bigcap_{N=1}^{\infty} \sigma(A_N, A_N + 1, \ldots) = \mathcal{T}.
\]

\( \square \)

**Theorem 3.6** (Kolmogorov’s zero-one law). If \( A_1, A_2, \ldots \) is an independent sequence of events, then the tail \( \sigma \)-field is trivial: if \( A \in \mathcal{T} \) then \( P(A) \) is either 0 or 1.

**Proof.** Fix \( A \in \mathcal{T} \). The plan is to prove that \( A \) is independent of itself, so \( P(A) = P^2(A) \).
For any \( n \), we have \( A \in \sigma(A_n, A_{n+1}, \ldots) \). So we consider one such \( n \). Lemma 3.4, applied to the array of independent events
\[
A_1, A_2, \ldots, A_{n-1} \\
A_n, A_{n+1}, \ldots
\]
says that the \( \sigma \)-fields \( \sigma(A_1, \ldots, A_{n-1}), \sigma(A_n, A_{n+1}, \ldots) \) are independent.

Unfortunately we need a somewhat stronger result about joint independence. We apply Proposition 3.5 to the array
\[
A_1 \\
A_2 \\
\vdots \\
A_{n-1} \\
A_n, A_{n+1}, \ldots
\]
of independent events. Then \( \sigma \)-fields \( \sigma(A_1), \ldots, \sigma(A_{n-1}), \sigma(A_n, A_{n+1}, \ldots) \) are independent.

Since \( A \in \mathcal{T} \subset \sigma(A_n, A_{n+1}, \ldots) \), we see that events \( A, A_1, A_2, \ldots, A_{n-1} \) are independent for every \( n \), i.e., \( A, A_1, A_2, \ldots, A_{n-1} \) is an infinite sequence of independent events.

Lemma 3.4, applied to two sequences of independent events
\[
\begin{array}{c}
A \\
A_1 \\
A_2 \\
\vdots \\
A_n \\
A_{n+1} \\
\ldots
\end{array}
\]
shows that \( \sigma(A) \) and \( \sigma(A_1, A_2, \ldots) \) are independent. In particular, for every \( B \in \sigma(A_1, A_2, \ldots) \) we have \( P(A \cap B) = P(A)P(B) \). But \( A \in \mathcal{T} \subset \sigma(A_1, A_2, \ldots) \), so we can choose \( B = A \). We get \( P(A) = P(A \cap A) = P(A)P(A) \).

\[ \square \]

**Example 3.4.** If \( A_n \) are independent events then \( P(\bigcap_{n} \bigcup_{k \geq n} A_k) \) is either 0 or 1. It is of interest to determine when each of the cases occurs.

**Corollary 3.7.** If \( A_n \) are independent events and \( \Omega_0 = \{ \omega : \frac{1}{n} \sum_{k} I_{A_n}(\omega) \text{ converges} \} \) then \( P(\Omega_0) \) is either 0 or 1. It is of interest to determine when each of the cases occurs. The celebrated law of large numbers (stated as Theorem 5.17 or as Theorem 8.4 below) says that \( P(\Omega_0) = 1 \) when the probabilities \( P(A_k) = p \) are all constant.

### 3. Borel-Cantelli Lemmas

**Theorem 3.8** (First Borel-Cantelli Lemma). If \( \sum_{n=1}^{\infty} P(A_n) < \infty \) then \( P(A_n \ i.o.) = 0 \).

**Example 3.5.** Suppose intervals \( A_n \subset (0,1] \) have lengths \( \lambda(A_n) = 1/n^2 \). Then the set of \( \omega \in \Omega \) for which \( \sum_{n} I_{A_n}(\omega) < \infty \) has Lebesgue measure 1, because with probability one only the finite number of \( I_{A_n}(\omega) \) is one, i.e. for almost every \( \omega \), this is a finite sum. In particular, \( I_{A_n} \to 0 \) for \( \omega \in \Omega_0 \) of probability one and \( P(A_n \ i.o.) = 0 \).

To appreciate this result better, consider a couple of situations for the case \( \lambda(A_n) = 1/n \). If \( A_n \) are adjacent intervals (wrapped around) of lengths \( 1/n \), then for every \( \omega \in [0,1] \) the sequence \( I_{A_n}(\omega) \) is 1 infinitely often, so \( \bigcap_{n} \bigcup_{k \geq n} A_k = \Omega \) and in particular \( P(A_n \ i.o.) = 1 \). You should contrast this with the case where \( A_n = [1 - 1/n, 1] \) where \( \bigcap_{n} \bigcup_{k \geq n} A_k = \{1\} \) so \( P(A_n \ i.o.) = 0 \).
3. Borel-Cantelli Lemmas

Proof. $0 \leq P(A_n \ i.o) = \lim_{n \to \infty} P(\bigcup_{k=n}^\infty A_k) \leq \lim_{n \to \infty} \sum_{k=n}^\infty P(A_k) = 0$. To prove the inequality used, recall that by continuity of $P$ and Exercise 1.2,

$$P(\bigcup_{k=n}^\infty A_k) = \lim_{m \to \infty} P(\bigcup_{k=n}^m A_k) \leq \lim_{m \to \infty} \sum_{k=n}^m P(A_k) = \sum_{k=n}^\infty P(A_k).$$

□

The following form of the lemma is sometimes more useful in the proofs.

Corollary 3.9. If $\sum_{n=1}^\infty P(A_n^c) < \infty$ then $P(A_n \text{ eventually}) = 1$.

Proof. $1 - P(A_n \text{ eventually}) = 1 - P(\bigcup_{n=1}^\infty \bigcap_{k>n} A_k) = P(\bigcap_{n=1}^\infty \bigcup_{k>n} A_k^c) = P(A_n^c \ i.o.)$ □

Theorem 3.10 (Second Borel-Cantelli Lemma). If $\sum_{n=1}^\infty P(A_n) = \infty$ and $A_n$ are independent events then $P(A_n \ i.o.) = 1$.

Proof. $P(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k) = 1 - P(\bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k^c)$ and $P(\bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k^c) \leq \sum_{n=1}^\infty P(\bigcap_{k=n}^\infty A_k^c)$.

Now it turns out that for every $n$, $P(\bigcap_{k=n}^\infty A_k^c) = 0$, since it is given as

$$\lim_{m \to \infty} P\left(\bigcap_{k=n}^m A_k^c\right) = \lim_{m \to \infty} \prod_{k=n}^m (1 - P(A_k)) \leq \lim_{m \to \infty} \exp\left(-\sum_{k=n}^m P(A_k)\right) = 0.$$

Here, we used the inequality $1 - x \leq e^{-x}$.

Corollary 3.11. If $A_1, A_2, \ldots$ are independent events then the following conditions are equivalent:

(i) $\sum_{n=1}^\infty P(A_n) < \infty$

(ii) $P(A_n \ i.o.) = 0$

Proof. (i)⇒(ii) is Borel-Cantelli-I. (ii)⇒(i) is by contrapositive and Borel-Cantelli-II: $\sum_{n=1}^\infty P(A_n) = \infty$ implies $P(A_n \ i.o.) = 1$ so $P(A_n \ i.o.) > 0$ □

Also their “negations” are equivalent:

Corollary 3.12. If $A_1, A_2, \ldots$ are independent events then the following conditions are equivalent:

(i') $\sum_{n=1}^\infty P(A_n) = \infty$

(ii') $P(A_n \ i.o.) = 1$
Exercise 3.1. Suppose $A, B, C$ are independent events. Prove directly from the definition that $(A \cup B), C$ is a pair of independent events. Similarly, show that events $(A \setminus B), C$ are independent.

Solution: Using exclusion-inclusion principle

\[
P((A \cup B) \cap C) = P((A \cap C) \cup (B \cap C)) =
\]
\[
P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) =
\]
\[
P(A)P(C) + P(B)P(C) - P(A)P(B)P(C) =
\]
\[
(P(A) + P(B) - P(A)P(B)) P(C) =
\]
\[
(P(A) + P(B) - P(A \cap B)) P(C) =
\]
\[
P(A \cup B)P(C)
\]

For the second proof, notice that $A \setminus B = A \cap B'$

Exercise 3.2 (Statistics\(^{2}\)). Consider $\Omega = [0, 1]$ with Lebesgue measure. Exhibit explicitly three independent events $A, B, C \subset [0, 1]$ with $\lambda(A) = \lambda(B) = \lambda(C) = 1/2$.

Solution: There are numerous answers, for example

\[
A = [0, 1/2], B = [0, 1/4] \cup [1/2, 3/4], C = [0, 1/8] \cup [1/4, 3/8] \cup [1/2, 5/8] \cup [3/4, 7/8]
\]

It is clear that $\lambda(A) = \lambda(B) = \lambda(C) = 1/2$, and that $\lambda(A \cap B) = \lambda([0, 1/4]) = 1/4$. Similarly $\lambda(A \cap C) = \lambda([0, 1/8] \cup [1/4, 3/8]) = 1/4$ and $\lambda(B \cap C) = \lambda([0, 1/8] \cup [1/2, 5/8]) = 1/4$. Finally that $\lambda(A \cap B \cap C) = \lambda([0, 1/8]) = 1/8$

Exercise 3.3. If $P(A_n) \geq \varepsilon > 0$ for all $n$, then $P(A_n \ i.o.) > 0$

Solution: Try proof by contradiction: Suppose $P(A_n \ i.o.) = 0$, i.e. $P(\bigcap_{n=1}^{\infty} A_n) = 0$. Then $P(\bigcup_{n=1}^{\infty} A_n^c) = 1$ so $\lim_{n \to \infty} P(A_n^c) = \lim_{n \to \infty} P(\bigcap_{k \geq n} A_k^c) = P(\bigcup_{n=1}^{\infty} A_n^c) = 1$. So $P(A_n) = 1 - P(A_n^c) \to 0$, contradicting the assumption that $P(A_n) \geq \varepsilon > 0$ for all $n$.

Exercise 3.4. Suppose $A_1, A_2, \ldots, A_n, \ldots$ are are independent events such that $P(A_n) = 1/2$. Show that

\[
\Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = 1.
\]

Solution: This can be handled using Borel-Cantelli Lemma, or directly by looking at the complement: $\Pr(\bigcap_{n=1}^{\infty} A_n^c) = \lim_{n \to \infty} \frac{1}{2^n} = 0$

Exercise 3.5. Suppose $A_1, A_2, \ldots, A_n, \ldots$ are independent events with probability $P(A_n) = n^{-\theta}$. Determine all $\theta \in \mathbb{R}$ such that $P(A_n \ i.o.) = 1$.

\(^{2}\)This allows you to simulate tosses of three coins using a single value from the random number generator that chooses uniformly numbers between 0 and 1.
Solution: By the zero-one law, \( P(A_n \text{ i.o.}) \) is either zero or one. Using Borel-Cantelli lemmas, if \( \sum_n P(A_n) = \infty \) then \( P(A_n \text{ i.o.}) = 1 \). On the other hand, if \( \sum_n P(A_n) < \infty \) then \( P(A_n \text{ i.o.}) = 0 \). Since \( \sum_n n^{-\theta} = \infty \) iff \( \theta \leq 1 \), the answer is \( P(A_n \text{ i.o.}) = 1 \) iff \( \theta \leq 1 \) and \( P(A_n \text{ i.o.}) = 0 \) iff \( \theta > 1 \).

Exercise 3.6. Suppose \( A_1, A_2, \ldots, A_n, \ldots \) are events with probability \( P(A_n) = 1/\log(n+1) \) so that by Corollary 3.11 we have \( P(A_n \text{ i.o.}) = 1 \). Use Borel-Cantelli Lemma to determine a subsequence \( n_k \to \infty \) such that events \( B_k = A_{n_k} \) satisfy \( P(B_k \text{ i.o.}) = 0 \). *(Hint: using base 2 for log might be more convenient for writing an example of explicit \( n_k \)).*

Solution: One possible answer is that \( P(A_{2^k} \text{ i. o.}) = 0 \), as \( P(A_{2^k}) = \frac{1}{\ln(2^{2^k} + 1)} \leq \frac{1}{\ln 2^{2^k}} = \frac{1}{k^2 \ln 2} \) so the series \( \sum_{k=1}^{\infty} P(A_{2^k}) \) converges.

Exercise 3.7. Suppose \( P \) is a probability measure on \( \mathbb{R}^2 \) with the cumulative distribution function \( F(x,y) \) that factors into a product of \( A(x)B(y) \). Prove that the \( \sigma \)-fields \( \mathcal{F} = \{ U \times \mathbb{R} : U \in B(\mathbb{R}) \} \) and \( \mathcal{G} = \{ \mathbb{R} \times U : U \in B(\mathbb{R}) \} \) are independent.

Here is a longer version of Exercise 3.1.

Exercise 3.8. Suppose \( A, B, C \) are independent events. Prove directly from the definition that their complements \( A^c, B^c, C^c \) are also independent events.

Solution: One way to prove this result is to apply a lemma below three times: first to \( A_1 = A, A_2 = B, A_3 = C \), then to \( A_1 = A, A_2 = C', A_3 = B \) and finally to \( A_1 = C', A_2 = B', A_3 = A \).

**Lemma** If \( A_1, A_2, A_3 \) are mutuallindependent then \( A_1, A_2, A_3^c \) are also independent

**Sketch of proof.** We only need to verify three identities: \( P(A_1 \cap A_2^c) = P(A_1) - P(A_1 \cap A_3) = P(A_1) - P(A_1)P(A_3) = P(A_1)(1 - P(A_3)) \) Then repeat the same with \( A_2 \) and with \( A_1 \cap A_2 \) in place of \( A_1 \).

Exercise 3.9. Suppose \( \{ A_n \} \) are independent events satisfying \( P(A_n) < 1 \). Show that

\[
P(\bigcup_{n=1}^{\infty} A_n) = 1 \text{ iff } P(A_n \text{ i.o.}) = 1
\]
3. Independence

Solution: \( P(A_n \ i.o.) = 1 \) means \( P(\cap_n \cup_{k \geq n} A_k) = 1 \). This implies that \( P(\cup_{k=1}^\infty A_k) \geq P(\cap_n \cup_{k \geq n} A_k) = 1 \).

Conversely, suppose that \( P(\cup_{k=1}^\infty A_k) = 1 \) and that \( P(A_n) < 1 \) for all \( n \). We first prove by induction that \( P(\cup_{k=n}^\infty A_k) = 1 \) for all \( n \).

**Proof.** Indeed, \( n = 1 \) is true. Suppose \( P(\cup_{k=n}^\infty A_k) = 1 \) for some \( n \geq 1 \). Then by inclusion-exclusion formula we have \( 1 = P(\cup_{k=n}^\infty A_k) = P(A_n) + P(\cup_{k=n+1}^\infty A_k) - P(A_n)P(\cup_{k=n+1}^\infty A_k) \). Since \( P(A_n) < 1 \), this shows that \( P(\cup_{k=n+1}^\infty A_k) = 1 \). \( \square \)

Once we know that \( P(\cup_{k=n}^\infty A_k) = 1 \) for all \( n \), we see that \( P(A_n \ i.o.) = P(\cap_n \cup_{k \geq n} A_k) = \lim_{n \to \infty} P(\cup_{k=n}^\infty A_k) = 1 \).

**Exercise 3.10.** Use the definitions to prove the claims from Example 3.3.

Solution: \( \{A_n \ i.o.\} = \cap_n \cup_{k \geq n} A_k \) is a tail event because \( \cap_{n=1}^\infty \cup_{k \geq n} A_k = \cap_{n=2}^\infty \cup_{k \geq n} A_k = \cap_{n=3}^\infty \cup_{k \geq n} A_k \ldots = \cap_{n=m}^\infty \cup_{k \geq n} A_k \in \sigma(A_m, A_{m+1}, \ldots) \) for every \( m \).

Similarly, \( \{ \text{all but a finite number of events } A_n \text{ hold} \} = \cup_n \cap_{k \geq n} A_k \) is a tail event because \( \cup_n \cap_{k \geq n} A_k = (\cap_{k \geq 1} A_k) \cup (\cap_{k \geq 2} A_k) \cup \cdots = \cap_{k \geq m} A_k \cup \cap_{k \geq m+1} A_k \cup \cdots \in \sigma(A_m, A_{m+1}, \ldots) \) for every \( m \).

(Or because this is a complement of a tail event!)

**Exercise 3.11.** Consider probability space \( (\Omega, \mathcal{F}, P) = ((0, 1), \mathcal{B}, \lambda) \) (the Lebesgue measure). Let \( A_n \) be "consecutive intervals" of length \( p_n \), "wrapped around" if needed. Thus \( A_1 = (0, p_1) \), \( A_2 = (p_1, p_1 + p_2) \), and so on. (The best way to imagine this is to think of \( \Omega \) as a circle, with one point removed, and \( A_n \) being "rotated" into the adjacent position next to \( A_{n-1} \).)

Borel-Cantelli Lemma gives a sufficient condition for \( P(A_n \ i.o.) = 0 \). Prove that \( P(A_n \ i.o.) = 1 \) iff \( \sum_n p_n = \infty \).

So \( P(A_n \ i.o.) \) is either 0 or 1 for "consecutively placed intervals", just like for independent events.

Solution: This is quite similar to Example 4.7.

**Exercise 3.12.** Let \( A_k \in \mathcal{F} \). Show that if \( P(\cap_n \cup_{k=n}^\infty A_k) = 1 \), then for every \( A \in \mathcal{F} \) of positive probability the series \( \sum_n P(A \cap A_n) \) diverges.

Solution: Lets proceed by contrapositive. Suppose there exists \( A \) with \( P(A) > 0 \), i.e. \( P(A^c) < 1 \) such that \( \sum_n P(A \cap A_n) < \infty \). Then by Borel Cantelli lemma, \( P(A \cap \cup_n \cup_{k=n}^\infty A_k) = 0 \). Therefore, \( P(\cup_n \cup_{k=n}^\infty A_k) = P(A \cap \cup_n \cup_{k=n}^\infty A_k) + P(A^c \cap \cup_n \cup_{k=n}^\infty A_k) \leq 0 + P(A^c) < 1 \).

**Exercise 3.13.** Let \( A_k \in \mathcal{F} \). Show that if for every \( A \in \mathcal{F} \) of positive probability the series \( \sum_n P(A \cap A_n) \) diverges, then \( P(\cap_n \cup_{k=n}^\infty A_k) = 1 \).

Solution: Lets proceed by contrapositive. Suppose \( P(\cap_n \cup_{k=n}^\infty A_k) < 1 \). Then \( A = (\cap_n \cup_{k=n}^\infty A_k)^c \cup \cup_n \cap_{k \geq n} A_k^c \) has positive probability. Now \( A \cap A_n = A_n \cap A = A_n \cap (A_1 \cap A_2 \cap \cdots \cap A_n \cap \cdots \cup (A_2 \cap A_3 \cap \cdots \cap A_n \cap \cdots \cup \cdots (A_n \cap A_{n+1} \cap A_{n+2} \cap \cdots)^c \cup \cdots (A_{n-1} \cap A_n \cap \cdots \cap A_{n+1} \cap A_{n+2} \cap \cdots)^c \cup \cdots (A_1 \cap A_2 \cap \cdots \cap A_n \cap \cdots)^c \cup \cdots) \) are disjoint, so \( \sum_n P(A \cap A_n) \leq 1 < \infty \).
Exercise 3.14. Suppose $A_1, A_2, \ldots$ are independent events. Consider

$$C = \left\{ \omega : \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} I_{A_k}(\omega) \text{ exists} \right\}.$$

Show that either $P(C) = 0$ or $P(C) = 1$.

(Solution: Hint: use zero-one law)

Exercise 3.15. Suppose $A_n$ is a sequence of events such that $\lim_{n \to \infty} P(A_n) = 0$ and $\sum_{n=1}^{\infty} P(A_n \cap A_{n+1}^c) < \infty$. Prove that $P(A_n \ i.o.) = 0$

Hints: [Resnik] gives the following hint: decompose $\bigcup_{j=n}^{\infty} A_j$. See also [Gut, Chapter 2, Theorem 18.7]. [Correct hint: decompose $\bigcup_{j=n}^{\infty} A_j^c$]

Solution: This is a theorem of Bandorff-Nielsen, who sketched the following argument in his paper.

Let $G = \bigcap_{n} \bigcup_{k \geq n} (A_k \cap A_{k+1}^c) = \{A_k \cap A_{k+1}^c \ i.o.\}$. By Borel Catnelli Lemma $P(G) = 0$. Next consider $F = \{A_k^c \ i.o.\} = \bigcap_{n} \bigcup_{k \geq n} A_k^c$. Then $P(F^c) = \lim_{n \to \infty} P(\bigcap_{k \geq n} A_k) \leq \lim_{n \to \infty} P(A_n) = 0$. So $P(F) = 1$.

Next, we look at $E = \{A_n \ i.o.\} = \bigcap_{n} \bigcup_{k \geq n} A_k$. Then $P(E) \leq P(F^c) + P(E \cap F)$. To end the proof, Since $P(F^c) = 0$ and $P(G) = 0$ it is enough to show that $E \cap F \subset G$. If $\omega \in E \cap F$ then, with fixed $n$, one can find $m \geq n$ such that $\omega \in A_m$. But $\omega \in F$, so from the definition of $F$, there exists some $k \geq m$ such that $\omega \in A_k^c$. Let's take first such $k$, which necessarily satisfies $k \geq m + 1$. This means $\omega \in A_{k-1} \cap A_k^c$. As $n$ was arbitrary, $\omega \in G$.

Exercise 3.16. Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Let $B = \{\omega : \sum_{n=1}^{\infty} I_{A_n}(\omega) < \infty\}$. Show that $P(B) = 1$.

Solution: $\sum_{n=1}^{\infty} I_{A_n}(\omega) = \infty$ iff $\{A_n \ i. o.\}$. By Borel-Cantelli Lemma, both events have probability zero.
Random variables

1. Measurable mappings

Suppose $\Omega$ and $E$ are two sets. Often $E = \mathbb{R}$ or $E = \mathbb{R}^d$.

Suppose $X : \Omega \rightarrow E$ i.e. $X$ is a function with domain $\Omega$ and target set $E$. Then $X$ induces a mapping

$$X^{-1} : 2^E \rightarrow 2^{\Omega}$$

defined by $X^{-1}(U) = \{ \omega \in \Omega : X(\omega) \in U \}$, where $U \subset E$.

**Proposition 4.1.** Properties of induced mapping:

(i) $X^{-1}(\emptyset) = \emptyset$, $X^{-1}(E) = \Omega$

(ii) $X^{-1}(U^c) = (X^{-1}(U))^c$

(iii) $X^{-1}(\bigcup_{t \in T} U_t) = \bigcup_{t \in T} X^{-1}(U_t)$

**Proof.** For (iii), $\omega \in X^{-1}(\bigcup_{t \in T} U_t)$ iff $\exists t \in T X(\omega) \in U_t$. □

**Corollary 4.2.** If $\mathcal{B}$ is a $\sigma$-field of subsets of $E$ then $X^{-1}(\mathcal{B})$ is a $\sigma$-field of subsets of $\Omega$.

**Proof.** This is based on the identities for inverse images under functions, see Proposition 4.1. □

**Definition 4.1.** A $\sigma$-field generated by $X$ is $\sigma(X) = X^{-1}(\mathcal{B})$.

Exercise 4.21 says that this is the smallest $\sigma$-field of subsets of $\Omega$ which makes $X$ measurable.

1.1. Random elements and random variables. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and $E$ is a set with distinguished $\sigma$-field $\mathcal{B}$. In most applications, $E$ is a separable complete metric space and $\mathcal{B}$ is the Borel $\sigma$-field which is generated by the countable collection of open balls.

**Definition 4.2.** In analysis, $X$ is called a measurable function if $X^{-1}(\mathcal{B}) \subset \mathcal{F}$. In probability, $X$ is then called a random element of $E$.
If we want to indicate the $\sigma$-fields, we will write $X : (\Omega, \mathcal{F}) \to (\mathbb{E}, \mathcal{B})$.

The most important special cases for us are $\mathbb{E} = \mathbb{R}$ and $\mathbb{E} = \mathbb{R}^d$. When $\mathbb{E} = \mathbb{R}$, we say that $X$ is a random variable. When $\mathbb{E} = \mathbb{R}^d$, we say that $X$ is a random vector or that $(X_1, \ldots, X_d)$ is a multivariate random variable. In such cases, measurability can be verified somewhat easier. To verify whether $X : \Omega \to \mathbb{R}$ is a random variable we only need to verify that the sets $A_x = \{ \omega : X(\omega) \leq x \}$ are in $\mathcal{F}$ for every real $x$.

Similarly, to verify whether $(X,Y) : \Omega \to \mathbb{R}^2$ is measurable, we only need to verify whether for all $x,y \in \mathbb{R}$ we have $\{ \omega : X(\omega) \leq x, Y(\omega) \leq y \}$ is in $\mathcal{F}$.

Somewhat more generally, we have the following.

**Proposition 4.3.** If $\sigma(\mathcal{A}) = \mathcal{B}_\mathbb{E}$ and $X^{-1}(\mathcal{A}) \subset \mathcal{F}$ then $X : (\Omega, \mathcal{F}) \to (\mathbb{E}, \mathcal{B})$ is measurable with respect to $(\mathcal{F}, \mathcal{B}_\mathbb{E})$.

**Proof.** Consider the set $\mathcal{U}$ of all sets $U \subset \mathbb{R}$ such that $X^{-1}(U) \in \mathcal{F}$. In view of Proposition 4.1, this is a sigma-field.

For $A \in \mathcal{A}$, the inverse image of the set $A$ is in $\mathcal{F}$, so $A \in \mathcal{U}$. Thus $\mathcal{A} \subset \mathcal{U}$, and the generated sigma field $\sigma(\mathcal{A}) = \mathcal{B}_\mathbb{E}$ is in $\mathcal{U}$. 

\[\square\]

**Remark 4.4.** The collection $X_1, \ldots, X_d$ of random variables (on the same probability space) defines random vector $(X_1, \ldots, X_d)$. (For $d = 2$, this is Exercise 4.20.)

We also remark that random elements of spaces of functions, such as $\mathbb{E} = C[0,1]$, the space of all continuous functions on $[0,1]$, or $\mathbb{E} = D[0,\infty)$, the space of right-continuous functions with left limits, are called stochastic processes rather than random functions. So we say ”Wiener process” or ”Poisson process”, rather than random continuous function, or random piecewise-linear function.

The following properties are often useful.

**Proposition 4.5.** Consider $\mathbb{R}$ or $\mathbb{R}^d$ with Borel sigma field.

- If $A \in \mathcal{F}$ then $I_A : \Omega \to \mathbb{R}$ is measurable.
- A continuous function $\mathbb{R} \to \mathbb{R}$ is measurable.
- A continuous function $\mathbb{R}^m \to \mathbb{R}^n$ is measurable.
- Composition of measurable transformations is measurable.
- Sum of two measurable functions is measurable.
- Product of two measurable functions is measurable.
- A pointwise limit of a sequence of measurable functions $\mathbb{R} \to \mathbb{R}$ is a measurable function.

For example, as a product of two measurable function $x \mapsto \sum_{a,b} e^{i(x \cdot a + y \cdot b)}$ is a measurable function $(\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$. For measurable $f,g$ the sum $f + g$ is measurable as a composition of $(f,g)\Omega \to \mathbb{R}^2$ with continuous function $(x,y) \mapsto x + y$.

1.2. Induced probability measures.

**Definition 4.3.** The distribution of a random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{E}, \mathcal{B})$ is a probability measure $\mu$ on $(\mathbb{E}, \mathcal{B})$ defined by

$$\mu(U) = P(X^{-1}(U))$$
Sometimes \( \mu \) is called an induced measure and some authors use notation \( Q = P \circ X^{-1} \). We will sometimes write \( L(X) = \mu \) and say that \( \mu \) is the law of \( X \).

If \( X \) is a random variable, then its distribution is uniquely determined by the corresponding cumulative distribution function

\[
F(x) = \mu((-\infty, x]) = P(\{\omega : X(\omega) \leq x\})
\]

In probability and statistics the latter is usually abbreviated to \( F(x) = P(X \leq x) \) but this abbreviated notation is just the shorthand for the right hand side of (4.1).

**Definition 4.4.** We say that random variables \( X, Y \), defined perhaps on different probability spaces, are equal in distribution, if they induce the same probability measure on \((\mathbb{R}, \mathcal{B})\).

In view of Proposition 2.19, this is equivalent to \( X, Y \) having the same cumulative distribution function.

**Example 4.1.** Consider \( \Omega = [0,1] \) with Lebesgue measure on Borel subsets. Let \( X(\omega) = \omega \) and \( Y(\omega) = 1 - \omega \). Then \( X, Y \) have the same distribution, which is easily checked by computing both cumulative distribution functions.

If \( X, Y \) are two random variables on the same probability space \((\Omega, \mathcal{F}, P)\) then the pair \((X, Y)\) is a measurable mapping \( \Omega \to \mathbb{R}^2 \). The joint distribution of random variables is just the induced measure on \((\mathbb{R}^2, \mathcal{B})\) and is uniquely determined by the joint cumulative distribution function

\[
F(x, y) = P(X \leq x, Y \leq y)
\]

(Note the abbreviated notation for \( P(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}) \)).

Taken together, pair \((X, Y)\) from Example 4.1 defines a probability measure on \((\mathbb{R}^2, \mathcal{B})\), and Exercise 4.9 asks for its joint cumulative distribution function. Cumulative distribution function of pairs of uniformly distributed random variables are called copulas and are useful in applications.

Our second proof of Theorem 4.10 below shows that one can find an infinite sequence of such functions on \([0,1]\), and furthermore, unlike in Example 4.1, the random variables can be independent.

(Omitted in 2019) If \( E = C[0,1] \) then a measurable mapping \( X : \Omega \to C[0,1] \) is called a stochastic process with continuous trajectories. The standard notation is \( X = \{X_t\}_{t \in [0,1]} \). The distribution of \( X \) is uniquely determined by the family of finite dimensional distributions

\[
F_{t_1, t_2, \ldots, t_k}(x_1, x_2, \ldots, x_k) = P(X_{t_1} \leq x_1, \ldots, X_{t_k} \leq k)
\]

that satisfy natural consistency conditions. The converse is not as simple here: consistent families of finite-dimensional distributions

\[
\{F_{t_1, t_2, \ldots, t_k}(x_1, x_2, \ldots, x_k) : 0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq 1\}
\]

define a probability measure on Borel sets of the product space \( E^{[0,1]} \) of all (including nonmeasurable) functions \([0,1] \to \mathbb{R}\) with pointwise convergence, see [Billingsley, Theorem 36.1] but not necessarily on Borel subsets of \( C[0,1] \). (In fact, \( C[0,1] \subset \mathbb{R}^{[0,1]} \) is not a Borel subset, see the discussion that follows [Billingsley, Theorem 36.3].) One way to ensure properties of trajectories that we need, is to construct the Wiener process and the Poisson process directly on some probability space \((\Omega, \mathcal{F}, P)\).

2. Random variables with prescribed distributions

This section is based on [Billingsley, Section 14] or [Durrett, Theorem 1.2.2].
4. Random variables

Theorem 4.6. If $F$ is a cumulative distribution function\(^1\), then there exists on some probability space a random variable $X$ for which $P(X \leq x) = F(x)$.

(Omitted in 2019)

**First proof.** Proposition 2.19 gives a probability measure $P$ on $(\mathbb{R}, \mathcal{B})$ such that $F(x) = P((0, x])$. Take $(\mathbb{R}, \mathcal{B}, P)$ for the probability space $(\Omega, \mathcal{F}, P)$. Define $X(\omega) = \omega$ (the identity mapping). Then $X$ has distribution $P$. \(\square\)

The second proof is independent of Proposition 2.19, and can be used to prove it.

**Second proof.** Let $\Omega = (0, 1)$ with Lebesgue measure $\lambda$ on Borel sigma-field. Since $F$ is non-decreasing right-continuous with limits $0, 1$, for $0 < u < 1$, the set $\{x : u \leq F(x)\}$ is\(^2\) a closed\(^3\) half-line\(^4\) of the form $[\varphi(u), \infty)$ and its complement is $\{x : u > F(x)\} = (-\infty, \varphi(u))$. This shows that for every real $x$, we have $\varphi(u) \leq x$ iff $F(x) \geq u$. This also defines the quantile function

\[
\varphi(u) = \inf\{x : u \leq F(x)\} = \sup\{x : F(x) < u\}
\]

Define $X(\omega) = \varphi(\omega)$. Then $\lambda(\{\omega : X(\omega) \leq x\}) = \lambda(\{\omega : F(\omega) \leq x\}) = \lambda((0, F(\omega)]) = F(x)$. \(\square\)

**Example 4.2** (Cauchy distribution). Cauchy distribution with density $\frac{1}{\pi (1 + x^2)}$ has cumulative distribution function $F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}$. For Cauchy random variable, the proof above gives formula $X(\omega) = \tan(\pi \omega - \pi/2)$.

**Corollary 4.7** (Proposition 2.19). If $F$ is a CDF then there exists a unique probability measure $P$ on the Borel sets of $\mathbb{R}$ such that $P((0, a]) = F(a)$.

**Proof.** Existence: Take Lebesgue measure on Borel sigma-field of $(0, 1)$, and $X$ as in the second proof above. Then $P$ is the induced probability measure. (Uniqueness follows from from $\pi - \lambda$ Theorem 2.13, see Proof of Proposition 2.19). \(\square\)

**Example 4.3.** Write $X = X_+ - X_-$, i.e.

\[
X_+(\omega) = \begin{cases} \frac{X(\omega)}{2} & \text{if } X(\omega) \geq 0 \\ 0 & \text{otherwise} \end{cases}
\]

\[
X_-(\omega) = (-X)_+ = \begin{cases} -X(\omega) & \text{if } X(\omega) \leq 0 \\ 0 & \text{otherwise} \end{cases}
\]

If $X$ has CDF $F(x)$, what are the CDFs of $X_+$ and $X_-$?

**Solution.**

\[
P(X_+ \leq x) = \begin{cases} P(X \leq x) & x \geq 0 \\ 0 & x < 0 \end{cases}
\]

So $F_+(x) = F(x)I_{[0,\infty)}(x)$. \(\square\)

---

\(^1\)See Definition 2.3
\(^2\)Can you see why isn’t it $\mathbb{R}$ or $\emptyset$?
\(^3\)Why?
\(^4\)Why?
2. Random variables with prescribed distributions

2.1. Independent random variables. The second proof of Theorem 4.6 lets us construct a finite or an infinite sequence $X_1 = \varphi_1(\omega), X_2 = \varphi_2(\omega), \ldots$ of random variables with prescribed distributions. However, this gives only very special measures on $\mathbb{R}^\infty$, see Exercise 4.22. We now consider another special construction that gives joint distributions that are of more interest.

**Definition 4.5.** Random variables $X_1, X_2, \ldots$ are independent if the generated $\sigma$-fields $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

In other words, $X_1, X_2, \ldots$ are independent if the events $X_1 \in U_1, X_2 \in U_2, \ldots$ are independent for any Borel sets $U_1, U_2, \ldots$.

**Proposition 4.8.** Random variables are independent, iff for every $n$ and every $x_1, \ldots, x_n \in \mathbb{R}$ the joint cumulative distribution function factors:

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = F_1(x_1) \cdots F_n(x_n)$$

**Proof.** Clearly, independence implies (4.4) - take $U_k = (-\infty, x_k]$.

(Omitted in 2019) To prove the converse, consider the families of $\pi$-systems $\mathcal{A}_k = \{X_k^{-1}((-\infty,x_k]), k = 1, 2, \ldots$. Then (4.4) implies that $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are independent, so by Corollary 3.3 the generated sigma fields are also independent.

---

**Example 4.4.** Consider discrete random variables $X = \sum_{j=1}^\infty x_j I_{A_j}, Y = \sum_{k=1}^\infty y_k B_k$ with disjoint sets. Then $X, Y$ are independent iff $\mathcal{A} = \{\emptyset, A_1, A_2, \ldots\}$ and $\mathcal{B} = \{\emptyset, B_1, B_2, \ldots\}$ are independent $\pi$-systems. Thus $X, Y$ are independent iff

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x, y \in \mathbb{R}$$

Similarly, discrete random variables $X, Y, Z$ are independent iff

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y)P(Z = z) \text{ for all } x, y, z \in \mathbb{R}$$

**Remark 4.9.** An important special case of discrete random variables are the simple random variables, which take only a finite number of values.

**Example 4.5.** Suppose $X_1, X_2, \ldots$ take only values 0, 1 and $p_k = P(X_k = 1), q_k = 1 - p_k$. Then $X_1, X_2, \ldots$ are independent iff

$$P(X_1 = \varepsilon_1, X_2 = \varepsilon_2, \ldots, X_n = \varepsilon_n) = \prod_{k=1}^n p_k^{\varepsilon_k} q_k^{1-\varepsilon_k}$$

for all $n$ and all choices of $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$.

Independence is often assumed in the theorems. So it is of some interest to make sure that such random variables exist.

**Theorem 4.10.** If $F_1, F_2, \ldots$ are cumulative distribution functions then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a sequence $X_1, X_2, \ldots$, of independent random variables such that $X_n$ has cumulative distribution function $F_n$. 
Sketch of First Proof. In this proof we take \( \Omega = \mathbb{R}^\infty \) with (infinite!) product measure\(^5 \) \( P = P_1 \otimes P_2 \otimes \ldots \) where \( P_k \) is the probability measure on \( \mathbb{R} \) with cumulative distribution function \( F_k \). For \( \omega = (\omega_1, \omega_2, \ldots) \in \mathbb{R}^\infty \) we define \( X_k(\omega) = \omega_k \). \( \square \)

Sketch of Second Proof. \(^6 \) The second proof is based on the idea that digits of \( \omega \in (0, 1) \) under Lebesque measure are independent and can be arranged into infinite number of sequences of (still independent) digits. This can be done in many ways, for example if \( \omega = .d_1d_2\ldots \) then we can rearrange its digits into

\[
\begin{array}{cccccc}
d_1 & d_2 & d_6 & d_7 & \ldots \\
d_3 & d_5 & d_8 & \vdots \\
d_4 & d_9 \\
d_{10} \\
\vdots
\end{array}
\]

splits \( \omega \in (0, 1) \) into the infinite sequence of numbers \( \omega_1 = .d_1d_2d_6d_7\ldots, \omega_2 = .d_3d_5d_8\ldots, \omega_3 = .d_4d_9\ldots, \) and so on.

We use \( \Omega = (0, 1) \) with Lebesgue measure \( \lambda \) and with binary digits function \( d_n : (0, 1] \rightarrow \{0, 1\} \).

We first note that random variables \( d_1, d_2, \ldots \) are independent. Indeed, as noted in the proof of Proposition B.1 we have \( \lambda(d_1 = \varepsilon_1, \ldots, d_m = \varepsilon_m) = 1/2^m \). By Example 4.5 this proves independence.

Next, we arrange all of these random variables into an infinite array \( d_{i,j} \). Then by Corollary 3.5 random variables \( U_i(\omega) = \sum_{j=1}^\infty d_{i,j}(\omega)/2^j \) are independent. On the other hand, \( \lambda(\omega : U_i(\omega) \leq x) = x \); this is easiest to see for diadic rational numbers\(^7 \) of the form \( x = k/2^n \).

Now take \( X_k = \varphi_k(U_k) \), where \( \varphi_k(u) \) is the quantile transform (4.3) of \( F_k \). \( \square \)

Definition 4.6. We say that \( X_1, X_2, \ldots \) are independent identically distributed (i. i. d.) random variables, if they are independent and have the same CDF.

Data collected from repeated runs of an experiment in statistics are modeled by i. i. d. random variables.

2.2. Elementary examples. The following is a repeat of one of the points in Proposition 4.5.

Proposition 4.11. If \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is measurable (say, continuous) and \( X_1, \ldots, X_d : \Omega \rightarrow \mathbb{R} \) are random variables on the probability space \( (\Omega, \mathcal{F}, P) \), then \( Y = f(X_1, \ldots, X_d) \) is a random vector.

Proof. If \( B \) is a Borel subset of \( \mathbb{R} \) then \( U = f^{-1}(B) \subset \mathbb{R}^d \) is a Borel subset of \( \mathbb{R}^d \). So \( Y^{-1}(B) = (X_1, \ldots, X_d)^{-1}(U) \in \mathcal{F} \). \( \square \)

Here are some examples of such functions:

Proposition 4.12 (Sum theorems). Suppose \( X_1, X_2, \ldots \) are independent and \( S = X_1 + X_2 + \cdots + X_n \).

\(^5\)For more details see [Billingsley, Theorem 20.4]. (This proof also answers Exercise 3.2.)

\(^6\)Observe that the diadic intervals \([0, k/2^n]\) with \( k, n \in \mathbb{N} \) form a \( \pi \)-system. that generates \( \mathcal{B} \)
3. Convergence of random variables

(i) If $X_1, \ldots, X_n$ are i. i. d. Bernoulli random variables, i.e., $P(X_j = 1) = p$, $P(X_j = 0) = 1 - p$, then $S$ is Binomial $\text{Bin}(n, p)$ (see Example 1.5)

(ii) If $X_1, X_2, \ldots$ are Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots$ then $S$ is Poisson with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$ (see Example 1.6)

(iii) If $X_1, X_2, \ldots$ are i. i. d. Normal $\text{N}(0,1)$ random variables (see Example 2.5) then $Y = X_1 + \cdots + X_n$ is normal with mean zero and variance $n$ (i.e., has same law as $\sqrt{n}Z$ for some $\text{N}(0,1)$ r.v. $Z$.)

Proof. Omitted$^8$

3. Convergence of random variables

The following definition requires us to know that $X_n - X$ is measurable. This is a consequence of Exercise 4.20.

Definition 4.7. A sequence of random variables converges in probability to a random variable $X$, abbreviated as $X_n \xrightarrow{p} X$, if for every $\varepsilon > 0$, we have

$$\lim_{n \to \infty} P(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}) = 0$$

We will use the abbreviated notation

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

It is also clear that it is enough to consider only rational $\varepsilon > 0$.

Example 4.6. On $\Omega = [0,1]$ consider $X_n = I_{[0,n/(2n+1)]}$. Then $X_n \xrightarrow{p} X$. In act, here convergence holds for every $\omega$. Exercise 4.12 shows that this does not have to be so. See also next example.

Example 4.7. Suppose $\Omega$ is a unit circle with (with probability measure from arclength, i.e. measure induced by $\theta \mapsto (\cos \theta, \sin \theta)$). Suppose $X_n = I_{\Theta_n}$ where $\Theta_n$ are consecutive arcs of length $1/n$ on the unit circle. Then $X_n \xrightarrow{p} 0$, but for every $\omega$ the sequence $X_n(\omega)$ does not converge.

Suppose $X_n, X$ are random variables on some probability space $(\Omega, \mathcal{F}, P)$. Then we have the following technical result.

Proposition 4.13.

$$\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\} \in \mathcal{F}$$

---

$^8$ These are “elementary” facts covered in undergraduate courses.
Proof. First we note that for a fixed $\varepsilon > 0$, the set $A_n = \{ \omega : |X_n(\omega) - X(\omega)| < \varepsilon \} \in \mathcal{F}$. This is a consequence of Exercise 4.20.

Next, we note that $A_{\varepsilon} = \{ \omega : \forall n \exists k > n |X_k(\omega) - X(\omega)| > \varepsilon \}$ is in $\mathcal{F}$. Indeed, $A_{\varepsilon} = \bigcap_n \bigcup_{k>n} A_k^c$.

Finally, we note that $\bigcap_{\varepsilon > 0} A_{\varepsilon} = \bigcap_n A_{1/n} \in \mathcal{F}$.

□

In view of the above proposition, the probability that a sequence of random variables converges to a random variable is well-defined.

**Definition 4.8.** A sequence of random variables converges with probability one if

$$P\left( \{ \omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega) \} \right) = 1$$

**Proposition 4.14.** If $X_n \to X$ with probability one, then $X_n \overset{P}{\to} X$

**Proof.** We need to observe that

$$P(\forall \varepsilon > 0 \exists n \forall n > N \{ \omega : |X_n - X| < \varepsilon \}) = 1 \text{ iff for every rational } \varepsilon > 0$$

$$P(\exists n \forall n > N |X_n - X| < \varepsilon) = P(\bigcup_n \bigcap_{n > N} |X_n - X| < \varepsilon) = 1$$

This is the same as

$$P\left( \bigcap_n \bigcup_{n > N} |X_n - X| > \varepsilon \right) = P(|X_n - X| > \varepsilon \text{ i.o.}) = 0$$

Now $P\left( \bigcap_n \bigcup_{n > N} |X_n - X| > \varepsilon \right) = \lim_{N \to \infty} P\left( \bigcup_{n > N} |X_n - X| > \varepsilon \right)$. So convergence with probability one is equivalent to

$$\forall \varepsilon > 0 \lim_{N \to \infty} P(\sup_{n > N} |X_n - X| > \varepsilon) = 0.$$ (4.5)

Of course, $P(|X_N - X| > \varepsilon) \leq P(\sup_{n > N} |X_n - X| > \varepsilon)$.

□

**Proposition 4.15.** Suppose $X_n \overset{P}{\to} X$. Then there exists a subsequence $n_k$ such that $X_{n_k} \to X$ with probability one

**Proof.** Choose positive $\varepsilon_k \to 0$, say $\varepsilon_k = 1/k$. Fix $k \in \mathbb{N}$. By convergence in probability, $\lim_{n \to \infty} P(|X_n - X| > \varepsilon_k) = 0$ so there is an integer $N$ such that for all $n \geq N$ we have $P(|X_n - X| > \varepsilon_k) < 1/2^k$. Choose one such $n$, and make sure that it is also large enough to satisfy $n > k$, too. Since this $n$ depends on $k$, call it $n(k)$ or $n_k$. This is the sub-sequence that we want. We now verify that $\lim_{k \to \infty} X_{n_k} = X$ with probability one. Since $\sum_k 1/2^k < \infty$, by the first Borel-Cantelli Lemma,

$$P(|X_{n_k} - X| > \varepsilon_k \text{ i.o.}) = 0$$

Therefore, for any $\varepsilon > 0$,

$$P(|X_{n_k} - X| > \varepsilon \text{ i.o.}) \leq P(|X_{n_k} - X| > \varepsilon_k \text{ i.o.}) = 0$$
Details: Choose $N_0$ such that $\varepsilon_{n_k} < \varepsilon$ for $k > N_0$. Then

$$\bigcap_{N=1}^{\infty} \bigcup_{k>N} \{|X_{n_k} - X| > \varepsilon\} \subset \bigcap_{N>N_0} \bigcup_{k>N} \{|X_{n_k} - X| > \varepsilon\} \subset \bigcap_{N>N_0} \bigcup_{k>N} \{|X_{n_k} - X| > \varepsilon_k\}$$

Since $P(\bigcap_{N>N_0} \bigcup_{k>N} \{|X_{n_k} - X| > \varepsilon_k\}) = 0$, we have $P(\bigcap_{N=1}^{\infty} \bigcup_{k>N} \{|X_{n_k} - X| > \varepsilon\}) = 0$.

**Remark 4.16.** Convergence in probability is a metric convergence, although it is too early to write down the metric, see Exercise 6.4s. Convergence with probability one is not a "metric convergence".

**Remark 4.17.** Suppose $X_n$ are random variables such that $X_n(\omega)$ converges for all $\omega \in \Omega$. Then the limit $X(\omega) := \lim_{n \to \infty} X_n(\omega)$ is a random variable. Somewhat more generally, if $\lim_{n \to \infty} X_n(\omega)$ converges with probability one, then $\Omega = \Omega_0 \cup \Omega_1$ decomposes into the two measurable sets where convergence holds and where convergence fails. Then with $X(\omega)$ defined arbitrarily (say as $0$) for the $\omega \in \Omega_0$, we can still claim that $X_n \to X$ with probability one. the limit can be used to define the random variable $X$.

**Proof.**

$$\{\omega : X(\omega) \leq x\} = \bigcap_{j \in \mathbb{N}} \bigcup_{n \in \mathbb{N}} \bigcap_{k>n} \{\omega : X_k(\omega) \leq x + 1/j\}.$$

Convergence in probability is a “metric convergence” (the metric cannot yet be written, but appears on page 143), so it has the following property.

**Proposition 4.18.** If every subsequence of $\{X_n\}$ has a $P$-convergent subsequence, then all the limits must be equal (with probability one) and $X_n$ converges in probability.

(Omitted in 2019) To solve Exercise 4.25 we use the same idea in a more concrete setting.

**Proof.** If different subsequences converge to say $X'$ and $X''$, then by choosing a subsequence that alternates between the two subsequences we can check that $P(|X' - X''| > \varepsilon) = 0$ for every $\varepsilon > 0$, so $X' = X''$ with probability one. Let's denote the common limit by $X$. To prove convergence, to the above $X$, we proceed by contradiction. Suppose that $X_n$ does not converge in probability. Then there exists a subsequence $\{X_{n_k}\}_{k \in \mathbb{N}}$ such that $P(|X_{n_k} - X| > \varepsilon) > \delta > 0$. This subsequence cannot have a further subsequence that would converge to $X$.

To see that this is quite useful, try solving Exercise 4.25 without using Proposition 4.18. (Yes, it can be done!)

Every convergent sequence of numbers is bounded. An analog of this involves a separate concept which is introduced in Exercise 4.14.

### 3.1. Convergence in distribution

The third type of convergence, the so called convergence in distribution, or sometimes weak convergence, is somewhat different, as it is really convergence of the induced probability measures, not of the random variables. This topic will appear in Chapter 9, but we can give the definition here:

**Definition 4.9.** We say that a sequence of $\mathbb{R}$-valued random variables $X_1, X_2, \ldots, X_n, \ldots$ with cumulative distribution functions $F_1, \ldots, F_n, \ldots$ converges in distribution to a random variable $Y$ with cumulative distribution function $F$, if $F_n(x) \to F(x)$ for all continuity points $x$ of $F$. 


Notation $X_n \xrightarrow{D} X$ is often used, but one has to keep in mind that this is the convergence of induced measures $\mu_n$, and that random variables themselves could come from different probability spaces.

(Omitted in 2019)

**Example 4.8** (Normal approximation to Binomial). Suppose $X_n$ is Bin$(n,p)$ and $Z$ is normal $N(0,1)$, see Example 2.5. In Chapter 11, we will show a theorem (Theorem 11.2) that will imply that

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z.$$ 

**Example 4.9** (Poisson approximation to Binomial). Suppose $X_n$ is binomial Bin($n,p = \lambda/n$). In Example 9.4 we will show that $X_n \xrightarrow{D} Y$ where $Y$ is Poiss($\lambda$).

Here is a more elementary example that can be worked out directly from the definition.

**Example 4.10** (extrema). Suppose $U_n$ are i.i.d. $U(0,1)$ and $X$ is exponential with parameter $\lambda = 1$, see Example 2.4.

Then $n \min\{U_1, U_2, \ldots, U_n\} \xrightarrow{D} X$.

**Proof.** Let $x > 0$ and take $n$ large enough so that $x/n < 1$. Then $F_n(x) := P(n \min\{U_1, U_2, \ldots, U_n\} \leq x) = 1 - P(\min\{U_1, U_2, \ldots, U_n\} \geq x/n) = 1 - P(U_1 \geq x/n \cap U_2 \geq x/n \cap \cdots \cap U_n \geq x/n) = 1 - (1 - x/n)^n \to 1 - e^{-x}$. This proves convergence for all $x > 0$. If $x \leq 0$ then $F_n(x) = 0$ converges to $F_X(x)$, too. \qed

Convergence in distribution is a metric convergence (of measures on $(\mathbb{R}, \mathcal{B})$, with Levy’s metric defined on page 143 and in Exercise 9.10.). It is a good exercise to check that if $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$. (This is Theorem 9.2 in the notes.)

**Example 4.11.** Suppose $\Omega = (0,1)$ with Lebesgue measure. Define $X_n(\omega) =\begin{cases} \omega & \text{if } n \text{ is even} \\ 1 - \omega & \text{if } n \text{ is odd} \end{cases}$

Then $X_n \xrightarrow{D} X_1$ because $F_n(x) = F_1(x)$. But $X_n$ cannot converge in probability, as it has two subsequence with two different limits $X_1$ and $X_2$.

An analog of Exercise 4.13 does not hold, as $X_n + Y_n$ is undefined, or ill-defined. (However there is a "substitute" in Theorem 9.4 later on.)

The relations between the three modes of convergence are complicated:

(i) Convergence with probability one implies convergence in probability. (Proposition 4.14)

(ii) Convergence in probability implies convergence in distribution. (Theorem 9.2, which could be proven here.)

(iii) Sequences convergent in probability have subsequences that converge with probability one. This is Proposition 4.15.

(iv) Sequences convergent in distribution can be redefined onto a common probability space on which they converge pointwise (hence almost surely). This is Theorem 9.6, which could have been proven here.

Later on, we will also need a fact from Exercise 4.25, so this exercise will be solved in class.
Measurability.

Exercise 4.1. Suppose that \( \varphi : (0, 1) \to \mathbb{R} \) is strictly increasing. Prove that \( \varphi \) is measurable with respect to the Borel sigma-fields.

Solution: We use Proposition 4.3 with \( \mathcal{A} \) consisting of all sets \((-\infty, b]\). For an increasing function the inverse image \( \varphi^{-1}(0, b] \) can only be of the form \((0, a]\) or \((0, a] \), the latter including \((0, 1) \) and \(\emptyset\).
Indeed, if \( b \) is above or below the range of \( \varphi \) then we get either \((0, 1) \) or \(\emptyset\). If \( b = \varphi(a) \) is in the range of \( \varphi \), then \( \varphi^{-1}(0, b] \) is \((0, a] \). If \( b \) is in the range of a jump at point \( a \), i.e. \( \varphi(a-) < b < \varphi(a) \) or \( \varphi(a) < b < \varphi(a+) \) then \( \varphi^{-1}(0, b] \) is \((0, a] \) in the first case and \((0, a] \) in the second.

Exercise 4.2. Suppose that \( \varphi : (0, 1) \to \mathbb{R} \) is continuous. Prove that \( \varphi \) is measurable with respect to the Borel sigma-fields.

Solution: We use Proposition 4.3 with \( \mathcal{A} \) consisting of all open sets. For a continuous function, \( \varphi^{-1}(U) \) is open for \( U \in \mathcal{A} \), it is in \( \mathcal{B} \).
(Solution of Exercise 4.3 provides additional details!)

Exercise 4.3. Prove one/some/all of the statements in Proposition 4.5.

Solution: Note that the number of properties in Proposition 4.5 increased from the initial number of points.
If \( A \in \mathcal{F} \) then \( I_A : \Omega \to \mathbb{R} \) is measurable.

Proof. This was done in class. Use the fact that \( \{ \omega : I_A(\omega) \leq a \} \) is either \( \emptyset \) (when \( a < 0 \)) or \( A^c \) (when \( 0 \leq a < 1 \)) or \( \Omega \) (when \( a \geq 1 \)) \( \square \)

A continuous function \( \mathbb{R}^m \to \mathbb{R}^n \) is measurable.

Proof. This is based on Proposition 4.3 used with \( \mathcal{A} = \) open sets, together with the following property of continuous functions: if \( f : \mathbb{R}^m \to \mathbb{R}^n \) is continuous and \( U \subset \mathbb{R}^n \) is an open set, then \( f^{-1}U \) is an open set in \( \mathbb{R}^m \).
To prove the later fact, look at a closed set \( V = U^c \) instead. Then \( f^{-1}(V) \) must be closed: if \( x_n \in f^{-1}(V) \) and \( x_n \to x \) then \( f(x_n) \in V \) and \( y = f(x) = \lim_{n \to \infty} f(x_n) \in V \), as \( V \) is closed. So \( x \in f^{-1}(V) \), i.e. \( f^{-1}(V) \) is closed.
Alternatively, start with the family of closed sets from the beginning, as they also generate \( \mathcal{B} \).\( \square \)
Solution: Composition of measurable transformations is measurable.

**Proof.** This is direct application of definition of measurability. See notes from class. □

Sum of two measurable functions is measurable.

**Proof.** $X + Y$ is a composition of a measurable function $\Omega \ni \omega \mapsto (X(\omega), Y(\omega)) \in \mathbb{R}^2$ with the continuous (hence measurable) function $\mathbb{R}^2 \to \mathbb{R}$ given by $(x, y) \mapsto x + y$. □

Product of two measurable functions is measurable.

**Proof.** This is similar to the sum: function $\mathbb{R}^2 \to \mathbb{R}$ given by $(x, y) \mapsto xy$ is continuous. □

Solution: A pointwise limit of a sequence of measurable functions $\mathbb{R} \to \mathbb{R}$ is a measurable function.

**Proof.**

$$
\{ \omega : \lim_{n \to \infty} X_n(\omega) \leq x \} = \{ \omega : \forall \epsilon > 0, \exists n \in \mathbb{N} \forall k > n X_k(\omega) < x + \epsilon \}
$$

$$
= \bigcap_{\epsilon > 0} \bigcup_{n=1}^{\infty} \bigcap_{k=n+1}^{\infty} \{ \omega : X_k(\omega) < x + \epsilon \}
$$

Cumulative distribution functions.

**Exercise 4.4.** Consider probability space $((0,1), \mathcal{B}, \lambda)$. Suppose $X : (0,1) \to \mathbb{R}$ is given by $X(\omega) = \ln(\omega)$. Find the CDF of $X$.

Solution: $X(\omega) \leq x$ iff $\omega \leq e^x$. So

$$
F(x) = \begin{cases} 
1 & x \geq 0 \\
\frac{1}{x} & x < 0 
\end{cases}
$$

**Exercise 4.5.** Suppose

$$
F(x) = \begin{cases} 
0 & x < 1 \\
1/2 & 1 \leq x < 2 \\
3/4 & 2 \leq x < 3 \\
1 & x \geq 3 
\end{cases}
$$

Compute $\varphi(u)$ from the proof of Theorem 4.10.

Solution:

$$
\varphi(u) = \begin{cases} 
1 & 0 < u \leq 1/2 \\
2 & 1/2 < u \leq 3/4 \\
3 & 3/4 < u < 1 
\end{cases}
$$

If you want another such exercise, determine $\varphi$ for the mixed-type CDF from Example 2.6 on page 19. (The answer explains how mixed-type random variables may occur for functions of random variables that have a density.)
Exercise 4.6. Suppose $X : \Omega \to \mathbb{R}$ has CDF $F$. Let $Y = X^2$. What is the CDF of $Y$?

Solution:

\[
G(y) = P(X^2 \leq y) = \begin{cases} 
0 & y < 0 \\
P(-\sqrt{y} \leq X \leq \sqrt{y}) & y \geq 0
\end{cases} \quad \begin{cases} 
0 & y < 0 \\
F(\sqrt{y}) - \lim_{x \uparrow -\sqrt{y}} F(x) & y \geq 0
\end{cases}
\]

Exercise 4.7. Suppose $X : \Omega \to \mathbb{R}$ has CDF $F$. Let $Y = X I_{|X| \leq M}$ be the truncation of r.v. $X$ at level $M$. What is the CDF of $Y$?

Solution: For a similar calculation, see Example 4.3 on page 40.

Exercise 4.8. Suppose $U$ is uniform on $(0, 1)$. Let $X = U^2$, $Y = U^3$. What is their joint CDF? (See (4.2).)

Solution: For $x, y \in (0, 1)$ we have

\[
F(x, y) = \lambda(u < \sqrt{x}, u \leq \sqrt{y}) = \min\{\sqrt{x}, \sqrt{y}\}
\]

Somewhat more generally, when we apply a construction from the second proof of Theorem 4.6 to construct a pair of random variables with prescribed CDFs $F, G$ then the joint PDF is $F(x, y) = \min\{F(x), G(y)\}$. This is related to the concept of copula, which is a fancy name for the multivariate CDF with uniform marginals. For uniform $U$, bivariate random variable $U, U$ has CDF $C(u, v) = \min\{u, v\}$. For independent $U, V$, the corresponding copula is $C(u, v) = uv$.

Exercise 4.9 (an example of copula). Find the joint cumulative distribution function for a random vector $(X, Y)$ defined in Example 4.1.

Exercise 4.10 (Statistics). Use the second proof of Theorem 4.6 to describe how to simulate exponential random variables (see Example 2.4) using a random number generator that produces an infinite sequence of uniform $U(0, 1)$ random variables.

Solution: With $F(x) = 1 - \exp(-\lambda x)$ (or 0) the set $\{x : F(x) \geq \omega\}$ is the interval with $\varphi(\omega) = \frac{-\ln(1 - \omega)}{\lambda}$. So given uniform r.v. $U$, we can take $Y = -\ln(1 - U)/\lambda$. Since $U$ and $1 - U$ have the same uniform CDF, we can instead take $Y = -\log(U)/\lambda$. (This is the answer you would get from google.)

Independence.

Exercise 4.11. Consider $\omega = [0, 1]$ with Lebesgue measure and the measure-preserving map $f$ defined in Exercise 2.3. Show that events $A := [0, 1/2]$, $B := f^{-1}(A)$ and $C := f^{-1}(B)$ are independent. (This is one of the possible answers to Exercise 3.2.)

Solution: $B = f^{-1}(A) = \{x \leq 1/2 : 2x \leq 1/2\} \cup \{x > 1/2 : 2x - 1 \leq 1/2\} = [0, 1/4] \cup (1/2, 3/4]$. Similarly, we can write out explicitly $f^{-1}(B)$. Compare solution of Exercise 3.2.
Convergence.

Exercise 4.12. Suppose random variables

\[ X_n = \begin{cases} 
  n & \text{with probability } p_n \\
  0 & \text{with probability } 1 - p_n 
\end{cases} \]

Prove that

(i) if \( p_n \to 0 \) then \( X_n \overset{P}{\to} 0 \).

(ii) if \( \sum p_n < \infty \) then \( X_n \to 0 \) with probability one.

(iii) if \( X_n \) are independent then \( X_n \to 0 \) with probability one iff and only if \( \sum p_n < \infty \)

Exercise 4.13. Prove that if \( X_n \overset{P}{\to} X \) and \( Y_n \overset{P}{\to} Y \) then \( X_n + Y_n \overset{P}{\to} X + Y \).

Solution: Hint: Use triangle inequality.

Exercise 4.14. Suppose \( X_n \overset{P}{\to} X \). Show that \( \{X_n\} \) is stochastically bounded (which is the same as the sequence of laws being tight, compare Exercise 1.11), i.e. for every \( \varepsilon > 0 \) there exists \( K > 0 \) such that for all \( n \) we have \( P(|X_n| > K) < \varepsilon \).

Exercise 4.15. Use the result from Exercise 4.14 to prove that if \( X_n \overset{P}{\to} X \) and \( Y_n \overset{P}{\to} Y \) then \( X_n Y_n \overset{P}{\to} XY \).

Exercise 4.16. Suppose \( U_1, U_2, \ldots, U_n, \ldots \) are independent identically distributed \( U(0, 1) \) random variables (i.e. with cumulative distribution function \( F(x) = x \) for \( 0 < x < 1 \), see Example 2.2 on page 17). Show that the sequence \( Z_n = U_1 U_2 \ldots U_n \) converges with probability one.

Solution: Hint: Use Borel-Cantelli Lemma. For another solution, see Exercise 8.2

Solution (II): For another solution see Exercise 8.2

Kolmogorov’s zero-one law suggests that the limit must be a constant, and we guess that it is 0.

Let \( \varepsilon > 0 \). We want to show that \( P(|Z_n| > \varepsilon \text{ i.o.}) = 0 \) i.e. \( \lim_{n \to \infty} P(\bigcup_{k \geq n} |Z_k| > \varepsilon) = 0 \).

If \( \varepsilon \geq 1 \) then \( P(|Z_n| > \varepsilon) = 0 \) so \( P(\bigcup_{k \geq n} |Z_k| > \varepsilon) \leq \sum_{k=n}^{\infty} P(|Z_k| > \varepsilon) = 0 \). If \( \varepsilon < 1 \) then since \( U_j \leq 1 \) we must have \( Z_n \leq \min_{k=1}^{n} U_k \). Thus \( P(|Z_n| > \varepsilon) = P(Z_n > \varepsilon) \leq P(\min_{k=1}^{n} U_k > \varepsilon) = P(\bigcap_{k=1}^{n} \{U_k > \varepsilon\}) = P(U > \varepsilon)^n = (1 - \varepsilon)^n \). Since \( \sum_{n=0}^{\infty} (1 - \varepsilon)^n < \infty \), by Borel-Cantelli lemma, \( \lim_{n \to \infty} P(\bigcup_{k \geq n} |Z_k| > \varepsilon) = 0 \).

Exercise 4.17. If \( X_n \leq Y_n \leq Z_n \) for all \( n \) and \( X_n \to X, Z_n \to X \) in probability then \( Y_n \to X \) in probability.
Solution (H): Inequality \( X - \varepsilon \leq Y_n \leq X + \varepsilon \) follows from \( Z_n \leq X + \varepsilon \) and \( X_n \geq X - \varepsilon \).

So \( \{Z_n \leq X + \varepsilon\} \cap \{X_n \geq X - \varepsilon\} \subseteq \{X - \varepsilon \leq Y_n \leq X + \varepsilon\} \). Thus

\[
P(|Y_n - X| \leq \varepsilon) \geq P(X_n \geq X - \varepsilon, Z_n \leq X + \varepsilon)
\]

Since if \( P(A) \leq P(B) \) then \( P(A^c) \geq P(B^c) \) we get

\[
P(|Y_n - X| > \varepsilon) \leq P((\{X_n \geq X - \varepsilon\} \cap \{Z_n \leq X + \varepsilon\})^c)
\]

\[
= P((\{X_n < X - \varepsilon\} \cup \{Z_n > X + \varepsilon\}) \leq P(X_n < X - \varepsilon) + P(Z_n > X + \varepsilon)
\]

\[
\leq P(|X_n - X| > \varepsilon) + P(|Z_n - X| > \varepsilon) \to 0
\]

**Exercise 4.18.** If \( X_n \to X \) in probability and \( X_n \geq 1 \) for all \( n \), then \( X \geq 1 \) (that is, \( P(X \geq 1) = 1 \) and \( \sqrt{X_n} \to \sqrt{X} \) in probability.

Solution: Of course, this is the special case of Exercise 4.25, so the point here is to find an easier proof.

Solution (H):

First, we need to show that \( X \geq 1 \). We do this by contradiction/contrapositive. Suppose that \( P(X \geq 1) < 1 \), i.e. \( P(X < 1) > 0 \). Then there is \( \varepsilon > 0, \delta > 0 \) such that \( P(X < 1 - \varepsilon) > \delta > 0 \). Indeed, \( P(X < 1) = P(\bigcup_n \{X < 1 - 1/n\}) = \lim_{n \to \infty} P(X < 1 - 1/n) \).

So we get a lower bound for \( P(|X_n - X| > \varepsilon) \geq P(X_n \geq 1, X < 1 - \varepsilon) = P(X < 1 - \varepsilon) > \delta \).

Thus we found \( \varepsilon > 0 \) such that \( P(|X_n - X| > \varepsilon) \) cannot converge to 0.

Next, we need to show that \( P(|\sqrt{X_n} - \sqrt{X}| > \varepsilon) \to 0 \). This can be done in many ways:

(i) We can use an inequality \( |\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|} \). This is not a well-known inequality, so one should prove it first. Applying this inequality, we get

\[
P(|\sqrt{X_n} - \sqrt{X}| > \varepsilon) \leq P(\sqrt{|X_n - X|} > \varepsilon) = P(|X_n - X| > \varepsilon^2) \to 0
\]

(This proof does not need the assumption that \( X_n \geq 1 \).)

(ii) We can use algebra:

\[
\sqrt{x} - \sqrt{y} = (\sqrt{x} - \sqrt{y})\frac{\sqrt{x} + \sqrt{y}}{\sqrt{x} + \sqrt{y}} = \frac{x - y}{\sqrt{x} + \sqrt{y}}
\]

For \( x, y \geq 1 \) we get \( |\sqrt{x} - \sqrt{y}| \leq |x - y| \). So \( P(|\sqrt{X_n} - \sqrt{X}| > \varepsilon) \leq P(|X_n - X| > \varepsilon) \)

\[a\]Without loss of generality, take \( x > y \). Then the inequality is equivalent to \( (\sqrt{x} - \sqrt{y})^2 \leq x - y \) or \(-2\sqrt{x} \sqrt{y} \leq -2y\), which holds true for \( x > y \).

**Exercise 4.19.** Give a formula for a subsequence \( n_k \) such that \( X_{n_k} \to 0 \) with probability one, if it is known that

(i) for all \( t > 0 \) we have \( P(|X_n| > t) \leq \frac{17}{t^2 \sqrt{n}} \)

Solution: Let just make choices based on the proof of Proposition 4.15: we want to choose \( n(k) \) such that \( \sum_k P(|X_{n_k}| > 1/k) < \infty \). We have \( P(|X_{n(k)}| > 1/k) \leq \frac{17k^2}{\sqrt{n(k)}} \), so the choice \( n(k) = k^8 \) accomplishes our goal. (Numerous other choices are possible. For example, proceeding as in Solution of part (iii) we can fix \( t > 0 \) and then choose \( n(k) = k^2 \log^4 k \).)
(ii) $X_n$ has density $\ln n \exp(-\ln x) = \frac{\ln n}{n^x}$ for $x \geq 0$. \(\text{corrected}\)

**Solution:** Since we are not asked to produce "the best answer", there are numerous choices here. All are based on the formula $P(X_n > t) = \int_t^\infty \frac{\ln n}{n^x} \, dx = \frac{1}{n^t}$.

Here is one choice: $P(X_n(k) > 1/k) = \frac{1}{\sqrt[n]{n(k)}}$. We want to choose $n(k)$ so that the series on the right converges. Clearly $n(k) = 2^{k^2}$ will work as then $\sqrt[n]{n(k)} = 2^k$.

Or you can replace $1/k$ by $1/\ln k$ and then request that $\ln k \sqrt[n]{n(k)} \sim k^2$. This gives $n(k) = [k^{2\ln k}]$ which is "a more optimal choice" as this sequence grows less rapidly than $2^{k^2}$.

(iii) for all $u > 0$ we have $P(|X_n| > u) \leq \frac{17}{u \log(n+1)}$

**Solution:** Notice that cases (i) and (iii) here are easier than the proof of Proposition 4.15: since we are given a rate, we can choose a sub-sequence that will work for all $\varepsilon$. Here, we can take $n(k) = 2^{k^2}$, say. Then For every $\varepsilon > 0$ we have $P(|X_n(k)| > \varepsilon \ i.o.) = 0$ as $\sum_k P(|X_n(k)| > \varepsilon) \leq \sum_k \frac{17}{2k^2 \log 2} < \infty$. Then $P(\exists \varepsilon > 0, \varepsilon \in \mathbb{Q}|X_n(k)| > \varepsilon \ i.o.) = 0$ i.e.

$$P(\forall \varepsilon > 0 \exists k > N |X_n(k)| \leq \varepsilon) = 1$$

### Additional Exercises

Some of these exercises appeared on past prelims!

**Exercise 4.20.** Suppose $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are two measurable functions (with respect to the Borel $\sigma$-field $\mathcal{B}(\mathbb{R})$). Prove that $(X, Y) : \Omega \to \mathbb{R}^2$ is measurable (with respect to the Borel $\sigma$-field $\mathcal{B}(\mathbb{R}^2)$). (Hint: Proposition 4.3.)

**Exercise 4.21.** Prove that $\sigma(X)$ as defined in the notes (as $X^{-1}(B)$) is in fact the smallest $\sigma$-field for which $X$ is measurable. (This is the definition of $\sigma(X)$ in [Billingsley].)

**Exercise 4.22.** Suppose $X,Y$ are random variables with cumulative distribution functions $F(x)$ and $G(y)$, constructed as in the second proof of Theorem 4.6. Find the joint cumulative distribution function of $X,Y$.

**Exercise 4.23** (Statistics). Suppose $X,Y$ are independent $N(0,1)$ random variables. Verify that $X^2 + Y^2$ is exponential. Hint: compute CDF using polar coordinates.

**Exercise 4.24.** Suppose that $X_1 \leq X_2 \leq \cdots \leq X_n \leq X_{n+1} \leq \cdots$. If $X_n \xrightarrow{p} X$, show that $X_n \to X$ with probability one.

**Exercise 4.25.** Suppose $f : \mathbb{R} \to \mathbb{R}$ is a continuous function and $X_n \xrightarrow{p} X$. Prove that $Y_n = f(X_n)$ converges in probability to $Y = f(X)$. Hint: an elementary proof relies on Proposition 4.18.

*(Solution: This exercise has many applications, so it should be solved in class)*
Solution: **First solution:** As in the proof of Proposition 4.15, we proceed by contradiction.
Suppose that \( X_n \xrightarrow{P} X \) but \( f(X_n) \) does not converge in probability to \( f(X) \). This means that there exists an \( \varepsilon_0 > 0 \) such that \( P(|f(X_n) - f(X)| > \varepsilon_0) \) does not converge to zero. Therefore, there exists \( \delta > 0 \) and a sequence of natural numbers \( n(k) \to \infty \) such that
\[
P(|f(X_{n(k)}) - f(X)| > \varepsilon_0) > \delta
\]
This contradicts the assumptions as follows: subsequence \( X_{n(k)} \xrightarrow{P} X \), so by Proposition 4.15 we can choose further subsequence \( n(k(j)) \) such that \( X_{n(k(j))} \to X \) with probability one as \( j \to \infty \). Since \( f \) is continuous function, this implies that \( f(X_{n(k(j)}) \to f(X) \) with probability one as \( j \to \infty \). Proposition 4.14 then implies that \( f(X_{n(k(j)}) \xrightarrow{P} f(X) \), so
\[
P(|f(X_{n(k(j)}) - f(X)| > \varepsilon_0) \to 0.
\]
But \( P(|f(X_{n(k(j)}) - f(X)| > \varepsilon_0) > \delta \), a contradiction.

Solution: **Second solution:** By Proposition 4.15, every subsequence \( \{X_{n_k} : k \in \mathbb{N}\} \) of \( \{X_n\} \) has a sub-subsequence \( \{X_{n_{k_j}} : j \in \mathbb{N}\} \) that converges with probability one. Therefore \( f(X_{n_{k_j}}) \to f(X) \) as \( j \to \infty \) with probability one. So this subsequence converges in probability (Proposition 4.14) and hence by Proposition 4.18 the original sequence converges.

Solution: **Third solution:** This solution relies on Exercise 4.14, and the fact that a continuous function is uniformly continuous on a compact set.
Fix \( \varepsilon, \varepsilon_1 > 0 \). Choose \( K \) such that \( P(|X_n| > K) < \varepsilon_1 \) for all \( n \), including \( n = \infty \). By uniform continuity of \( f \) on \([-K,K]\), there exists \( \delta > 0 \) such that if \( |x - y| < \delta \) then \( |f(x) - f(y)| < \varepsilon \). Then
\[
P(|f(X_n) - f(X)| > \varepsilon)
\leq P(|f(X_n) - f(X)| > \varepsilon, |X_n| \leq K, |X| \leq K) + P(|X_n| > K \cup |X| > K)
\leq P(|X_n - X| > \delta) + 2\varepsilon_1
\]
Since \( P(|X_n - X| > \delta) \to 0 \) and \( \varepsilon_1 \) is arbitrary, this shows that \( P(|f(X_n) - f(X)| > \varepsilon) \to 0 \).

The following generalization of Exercise 4.25 can be proved by a slight elaboration on the same techniques.

**Exercise 4.26.** Suppose \( f : \mathbb{R}^2 \to \mathbb{R} \) is a continuous function and \( X_n \xrightarrow{P} X \), \( Y_n \xrightarrow{P} Y \). Prove that \( Z_n = f(X_n, Y_n) \) converges in probability to \( Z = f(X, Y) \).

**Exercise 4.27.** Suppose \( X_n \xrightarrow{P} X \) and \( X_n \) are independent. Show that there is \( a \in \mathbb{R} \) such that the cumulative distribution of \( X \) is \( F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases} \).
Solution: Here is a sketch of one solution: For each real $x$ the event $X \leq x$ must be a tail event, so by Kolmogorov’s zero-one law, $P(X \leq x)$ is either zero or one. For large $x$, we must have $P(X \leq x) = 1$. Let $U = \{ x \in \mathbb{R} : P(X \leq x) = 1 \}$. Then set $U$ is bounded from below, as $P(X \leq x) = 0$ for $x \to -\infty$. Take $a = \inf U$ to get your number. (Other more direct solutions are possible.)

\*This requires proof!}
Simple random variables

Printed: March 15, 2019
This section is based on [Billingsley, Section 5].

Definition 5.1. A random variable $X$ is a simple random variable if it has a finite range. If the range $X(\Omega)$ of $X$ is $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$ (distinct real numbers), then

$$X = \sum_{j=1}^{n} x_j I_{A_j},$$

where $A_j = X^{-1}(\{x_j\}) \in \mathcal{F}$. Note that if $x_j$ are distinct then $A_j$ are disjoint, and that $\bigcup_{j=1}^{n} A_j = \Omega$.

Theorem 5.1. Let $X_1, \ldots, X_n$ be simple random variables. A simple random variable $Y$ is $\sigma(X_1, \ldots, X_n)$-measurable if and only if there exists $f: \mathbb{R}^n \rightarrow \mathbb{R}$ such that $Y = f(X_1, \ldots, X_n)$.

Proof. If $Y = f(X_1, \ldots, X_n)$ then $Y^{-1}(y) = \{\omega: (X_1(\omega), \ldots, X_n(\omega)) \in f^{-1}(\{y\})\}$. Of course, $f^{-1}(\{y\})$ could be a non-measurable set. But its intersection with a finite set $F_1 \times F_2 \times \cdots \times F_n$ is measurable. So $Y^{-1}(\{y\})$ is an inverse image of a measurable set in a measurable mapping $(X_1, \ldots, X_n): \Omega \rightarrow \mathbb{R}^n$ (compare Exercise 4.20).

Suppose now that $Y$ is $\sigma(X_1, \ldots, X_n)$. Denote by $y_1, \ldots, y_r$ its distinct values. Then there exists a set $U_i \subset \mathbb{R}^n$ such that

$$\{\omega: Y(\omega) = y_i\} = \{\omega: (X_1(\omega), \ldots, X_n(\omega)) \in U_i\}$$

Take $f = \sum y_j I_{U_j}$. (The sets $U_j$ are not disjoint, but their intersections with the range of $(X_1, \ldots, X_n)$ are disjoint.)

The importance of simple random variables lies in their usefulness for approximations.

Theorem 5.2. If $X: \Omega \rightarrow [0, \infty)$ is a nonnegative random variable then there exist a sequence of simple random variables $X_1 \leq X_2 \leq X_3 \leq \cdots \leq X_n \leq \cdots$ such that $X(\omega) = \lim_{n \rightarrow \infty} X_n(\omega)$. 

(Omitted in 2019)
Remark 5.3. It is easy to produce good approximations on the sets of large probability,

\[ \sum_{k=1}^{n^2} \frac{k-1}{n} I_{\frac{k-1}{n} \leq X < \frac{k}{n}} \to X, \]

or discrete uniform approximations from below on entire \( \Omega \). For the latter, take

\[ X_n = \sum_{k=1}^{\infty} \frac{k-1}{n} I_{\frac{k-1}{n} \leq X < \frac{k}{n}}. \]

Then \( X - 1/n \leq X_n \leq X \).

For the proof, we want to use only finite number of values, work on all \( \Omega \), and be sure that the approximation is also increasing so that \( X_n \uparrow X \).

Proof. \(^1\) We find an appropriate function \( \varphi_n(x) \) and take as \( X_n \) the value of \( \varphi_n(X) \). Here is one such function:

\[ (5.2) \quad X_n := nI_{X \geq n/2^n} + \sum_{k=1}^{2^n} \frac{k-1}{n^2} I_{\frac{k-1}{n^2} \leq X < \frac{k}{n^2}} \uparrow X. \]

See Fig 1. \( \square \)

1. Expected value

A simple random variable of the form (5.1) is assigned expected value

\[ (5.3) \quad \mathbb{E}[X] = \sum_{j=1}^{n} x_j P(A_j) \]

Remark 5.4. A special case of (5.3) is

\[ \mathbb{E}[I_A] = P(A). \]

In particular, if \( \mathbb{E}[I_A] = 0 \) then \( P(A) = 0 \).

Remark 5.5. Note that if \( \Omega = [0,1] \) and \( A_j \) are intervals, then \( E(X) = \int_{0}^{1} X(\omega)d\omega \), defined as the Riemann integral.

Example 5.1. Suppose \( X \) is Binomial \( Bin(n,p) \), see Example 1.5 on page 7. Then \( x_j = j \) and \( P(A_j) = \binom{n}{j} p^j (1-p)^{n-j} \) with \( j = 0, 1, \ldots, n \), so

\[ \mathbb{E}[X] = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \]

\[ = np(p + 1 - p)^{n-1} = np. \]

\(^1\)This proof is repeated at the beginning of next chapter!
It is clear that if $X$ is simple and $f: \mathbb{R} \to \mathbb{R}$ is an arbitrary function then $Y = f(X)$ is simple, and that
\begin{equation}
\mathbb{E}[Y] = \sum_j f(x_j) P(A_j).
\end{equation}
In particular, the moments of $X$ are
\[ m_k = \mathbb{E}[X^k] = \sum_j x_j^k P(A_j). \]

**Proposition 5.6.** For simple random variables the expected value has the following properties:
- **linearity:**
  \begin{equation}
  \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]
  \end{equation}
  and more generally, $\mathbb{E}[aX + Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$.
- **monotonicity:** if $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- 
\begin{equation}
|\mathbb{E}[X]| \leq \mathbb{E}|X|.
\end{equation}
This implies $|\mathbb{E}[X - Y]| \leq \mathbb{E}|X - Y|$.

**Proof.** If $X = \sum_j x_j I_{A_j}$ and $Y = \sum_k y_k I_{B_k}$ then $X + Y = \sum_{j,k} (x_j + y_k) I_{A_j \cap B_k}$. Thus $E(X + Y) = \sum_{j,k} (x_j + y_k) P(A_j \cap B_k) = \sum_j x_j \sum_k P(A_j \cap B_k) + \sum_k y_k \sum_j P(A_j \cap B_k)$. This gives linearity (5.5).

Expected value also preserves order: if $X \geq 0$ for all $\omega$ then $E(X) \geq 0$. Thus if $X \leq Y$ (i.e. $Y - X \geq 0$) then $E(X) \leq E(Y)$.

Since $X \leq |X|$, this gives $E(X) \leq E|X|$. Since $-X$ satisfies this, too, we get (5.6).

**Definition 5.2.** We will say that $\{X_n\}$ is uniformly bounded if there exists a real number $K$ such that $|X_n(\omega)| \leq K$ for all $\omega \in \Omega$ and all $n$.

We say that $\{X_n\}$ is uniformly bounded with probability one if there exists a real number $K$ such that $P(|X_n(\omega)| \leq K) = 1$ for all $n$.

We say that $\{X_n\}$ is stochastically bounded, if for every $\varepsilon > 0$ there exists $K \in \mathbb{R}$ such that $P(|X_n| > K) < \varepsilon$ for all $n$.

**Example 5.2.** Here are some examples that illustrate the definition

(i) A sequence $X_n = \pm 1$ is of course uniformly bounded.

(ii) If $P(A_n) = 0$ and $A_n \neq \emptyset$ and $B \in \mathcal{F}$ then random variables $X_n = nI_{A_n} + I_{B \cap A_n} - I_{A_n \cap B^c}$ are uniformly bounded with probability one, but not uniformly bounded, as the values of $X_n$ are generically $\{-1, 1, \emptyset\}$.

(iii) A sequence of, say independent, Poisson random variables $X_n$ with the same distribution is stochastically bounded but not bounded.

**Theorem 5.7.** Suppose that $X_n, X$ are simple random variables such that $X_n \xrightarrow{P} X$. If $\{X_n\}$ is uniformly bounded, then $\mathbb{E}[X] = \lim_{n \to \infty} \mathbb{E}[X_n]$. 


Proof. Suppose \(|X_n| \leq K\). Since \(X\) is simple, we can increase \(K\) to ensure also \(|X| \leq K\).

If \(A_n = \{ \omega : |X - X_n| \geq \varepsilon \}\) then

\[
|X(\omega) - X_n(\omega)| \leq 2K I_{A_n} + \varepsilon I_{A_n}
\]

Thus \(E|X - X_n| \leq 2KP(|X_n - X| \geq \varepsilon) + \varepsilon \rightarrow \varepsilon\). Inequality (5.6) ends the proof. \(\Box\)

Example 5.3. Suppose \(P(X_n = 0) = (n - 1)/n\) and \(P(X_n = (-1)^n) = 1/n\). Then \(X_n \rightarrow 0\) but \(E(X_n) = (-1)^n\) does not converge. This contradicts Theorem 5.7, doesn’t it?

Remark 5.8. Suppose \(X \geq 0\) is arbitrary, and \(X_n \uparrow X\) are simple random variables from Theorem 5.1. Then \(E(X_n) \leq E(X_{n+1})\) so \(\lim_{n \rightarrow \infty} E(X_n)\) exists, perhaps as \(\infty\). Furthermore, if \(X\) is simple, then by Theorem 5.7, \(\lim_{n \rightarrow \infty} E(X_n)\) is just \(E(X)\). This suggests that we can try to define \(E(X)\) by this limit. (It would be nice to know that any other sequence \(X_n \uparrow X\) will give the same answer!)

It is tempting to compute by this technique an answer that we know from somewhere else. But anything more complicated than Exercises 5.5 or 5.6 seems to require too much work.

Definition 5.3. The variance of a simple random variable \(X\) is

\[
(5.7) \quad \text{Var}(X) = E(X - m)^2 = E(X^2) - m^2
\]

where \(m = E(X)\).

The mean and variance of a linear transformation \(Y = aX + B\) of \(X\) are \(E(Y) = aE(X) + b\), \(\text{Var}(Y) = a^2 \text{Var}(X)\).

Somewhat more generally,

\[
\text{Var}(aX + bY + c) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)
\]

where, trivially,

\[
\text{Cov}(X, Y) = \frac{1}{2} (\text{Var}(X + Y) - \text{Var}(X) - \text{Var}(Y))
\]

This gives

\[
\text{Cov}(X, Y) = E((X - m_X)(Y - m_Y)) = E(XY) - E(X)E(Y)
\]

2. Expected values and independence

If \(X_1, \ldots, X_n\) are independent then

\[
(5.8) \quad E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n)
\]

It is enough to verify this for two independent random variables. If \(X = \sum x_j I_{A_j}\) and \(Y = \sum y_k I_{B_k}\) then \(XY = \sum_{j,k} x_j y_k I_{A_j \cap B_k}\). Thus \(E(XY) = \sum_{j,k} x_j y_k P(A_j \cap B_k) = \sum_j x_j P(A_j) \sum_k y_k P(B_k)\). Thus \(E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n)\) and inductively we can pull one factor at a time.

For independent \(X_1, \ldots, X_n\) we have

\[
\text{Var}(X_1 + \ldots + X_n) = \text{Var}(X_1) + \ldots + \text{Var}(X_n)
\]

Again, we verify this for the sum of two independent variables \(X, Y\). Replacing \(X\) by \(X - m\) if needed, without loss of generality we may assume \(E(X) = E(Y) = 0\). Then \(\text{Var}(X + Y) = E(X^2 + Y^2 + 2XY) = E(X^2) + E(Y^2) = \text{Var}(X) + \text{Var}(Y)\).
2.1. Tail integration formula. If $X \geq 0$ then

\[ E(X) = \int_0^\infty P(X > x)dx = \int_0^\infty P(X \geq x)dx \]

**Proof.** For simple random variables this is just a picture. \(\square\)

**Remark 5.9.** Remark 5.8 suggested a definition of $E(X)$ that was hard to use in specific examples. Here is another approach. The function $x \mapsto P(X \geq x)$ is decreasing, so it is Riemann-integrable, \(\int_0^T P(X \geq x)dx\) exists. Furthermore, the integral is an increasing (or at least non-decreasing) function of its upper limit $T$, so for $X \geq 0$ it is tempting to define $\mathbb{E}[X]$ as the (possibly infinite) improper Riemann integral (5.9).

3. Inequalities

3.1. Markov inequality. Markov’s inequality says the following.

**Proposition 5.10.** For non-negative simple r.v. $X$ and $\alpha > 0$, we have

\[ P(X \geq \alpha) \leq \frac{1}{\alpha}E(X) \]

**Proof.** This follows from (5.9), as $E(X) \geq \int_0^\alpha P(X \geq x)dx \geq \int_0^\alpha P(X \geq \alpha)dx$. \(\square\)

From Remark 5.4 we see that one can have $\mathbb{E}[|X|] = 0$ even if $X = I_A$. But we have the following.

**Corollary 5.11.** If $\mathbb{E}[|X|] = 0$ then $P(X = 0) = 1$.

**Proof.** By (5.10) applied to non-negative random variable $|X|$, for every $\alpha > 0$, we have $0 \leq P(|X| > \alpha) \leq P(X \geq \alpha) = 0$. So $P(|X| > 0) = P(\bigcup_n |X| > 1/n) = \lim_n \to \infty P(|X| > 1/n) = 0$.

This implies Chebyshev’s inequality.

**Proposition 5.12.** For $\alpha > 0$,

\[ P(|X - m| \geq \alpha) \leq \frac{Var(X)}{\alpha^2} \]

**Corollary 5.13.** If $X$ is a simple random variable such that $Var(X) = 0$ and $\mathbb{E}(X) = m$ then $X = m$ with probability one.

**Proof.** Write $X = \sum x_j I_{A_j}$ and let $\alpha = \min\{|x_j - m| : x_j - m \neq 0\}$ be the smallest. Note that $\alpha > 0$. By (5.11),

\[ P(X - m \neq 0) = \sum_{j : x_j - m \neq 0} P(A_j) = P(|X - m| \geq \alpha) = 0 \]

Exercise 5.9 is another application of (5.10).
3.1.1. Jensen, Hölder. Recall that $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function\footnote{A sufficient condition is $\varphi''(x) \geq 0$.} if $\varphi(px + (1-p)y) \leq p\varphi(x) + (1-p)\varphi(y)$. Inductively, $\varphi(\sum x_j p_j) \leq \sum x_j \varphi(p_j)$. This gives Jensen’s inequality
\begin{equation}
(5.14) \quad \varphi(E(X)) \leq E(\varphi(X))
\end{equation}

Similarly, if $\varphi : \mathbb{R}^d \to \mathbb{R}$ is convex, then
$$
\varphi(EX_1, EX_2, \ldots, EX_d) \leq E(\varphi(X_1, X_2, \ldots, X_d))
$$

Special cases are $|E(X)| \leq E|X|$, $E(X)^2 \leq E(X^2)$, $\exp(E(X)) \leq E(\exp X)$, $E\ln X \geq \ln E(X)$

In particular, $E(|X|) \leq \sqrt{E(X^2)}$. More generally, we have Lyapunov’s inequality: if $\alpha \leq \beta$ then
\begin{equation}
(5.13) \quad E^{1/\alpha}(|X|^{\alpha}) \leq E^{1/\beta}(|X|^{\beta})
\end{equation}

Indeed, with $p = \beta/\alpha \geq 1$ function $\varphi(x) = |x|^p$ is convex\footnote{$f''(x) = p(p-1)x^{p-2} > 0$ for $x > 0$ and $p > 1$.}. Write $|X|^\beta = (|X|^{\alpha})^p = \varphi(|X|^{\alpha})$. Then by Jensen’s inequality,
$$
(E(|X|^{\alpha}))^{\beta/\alpha} \leq E|X|^{\beta}
$$

Another important inequality is Cauchy-Schwarz inequality
\begin{equation}
(5.14) \quad |E(XY)| \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}
\end{equation}

Proof #1. The simplest proof is to consider the quadratic polynomial in variable $t$ defined by $p(t) = E(Y + tX)^2$. (Without loss of generality we may assume that $E(Y^2) \neq 0$.) Since $p(t) \geq 0$ and $p(t) = E(Y^2) + 2tE(XY) + t^2E(X^2)$ we have $(E(XY))^2 \leq E(X^2)E(Y^2)$, as the quadratic polynomial $p(t)$ cannot have two real roots. \hfill $\Box$

Proof #2. Here is a proof using Jensen’s inequality: The function $(x, y) \to -\sqrt{x}\sqrt{y}$ is convex on $[0, \infty) \times [0, \infty)$. We apply Jensen’s inequality to non-negative random variables $X^2$ and $Y^2$. $E\sqrt{X^2}\sqrt{Y^2} \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}$ \hfill $\Box$

Proof #3. Here is a proof based on the elementary inequality $ab \leq a^2/2 + b^2/2$.

By homogeneity we may assume that $E(X^2) = E(Y^2) = 1$. Then we apply the elementary inequality with $a = x_i \sqrt{P(A_i \cap B_j)}$ and $b = y_j \sqrt{P(A_i \cap B_j)}$. We get
$$
E(XY) = \sum_{i,j} x_i y_j P(A_i \cap B_j) \leq \sum_{i,j} \frac{1}{2} x_i^2 P(A_i \cap B_j) + \sum_{i,j} \frac{1}{2} y_j^2 P(A_i \cap B_j) = 1
$$
3.1.2. *The correlation coefficient.* The correlation coefficient between non-degenerate random variables $X, Y$ is a real number defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Applying Cauchy-Schwartz inequality (5.14) to the product $X - m_X(Y - m_Y)$ we see that $|\rho| \leq 1$.

**Proposition 5.14.** If $\rho^2 = 1$ then, with probability one, $Y$ is a linear function of $X$.

**Proof.** Without loss of generality, we may assume $E(X) = E(Y) = 0$. Inspecting proof #1 of Cauchy-Schwartz, we see that polynomial $p(t)$ there has a minimum at $t_0 = -\frac{E(XY)}{E(X^2)} = -\rho\sqrt{\frac{E(Y^2)}{E(X^2)}}$. Inserting this into the definition, we see that

$$E(Y + tX)^2 \geq (1 - \rho^2)E(Y^2),$$

and that at $t_0$ we have equality. In particular, if $\rho^2 = 1$ then $E(Y + t_0X)^2 = 0$. So by Corollary 5.13, $|P(Y = t_0X) = 1|.$

\[\square\]

4. *$L_p$-norms*

For $p \geq 1$, define the $L_p$-norm of $X$ as

$$\|X\|_p = \sqrt[p]{\mathbb{E}(|X|^p)}$$

Lyapunov’s inequality says that of $p_1 \leq p_2$ then $\|X\|_{p_1} \leq \|X\|_{p_2}$. In particular, $\|X\|_1 \leq \|X\|_2$.

The Cauchy-Schwarz inequality can be stated concisely as

$$|E(XY)| \leq \|X\|_2 \|Y\|_2$$

It is clear that $\|\alpha X\|_p = |\alpha|\|X\|_p$ and that $\|X\|_p \geq 0$ is zero only if $X = 0$ (with probability one). What is less obvious is that this is indeed a norm in the vector space of all simple random variables. For this, we need a triangle inequality.

**Theorem 5.15** (*Minkowski’s inequality*).

(5.15) \[\|X + Y\|_p \leq \|X\|_p + \|Y\|_p\]

**Proof of Minkowski’s inequality for $p = 1$.** Using triangle inequality and monotonicity of expectation, we have

$$\|X + Y\|_1 = E|X + Y| \leq E(|X|) + E(|Y|) = \|X\|_1 + \|Y\|_1$$

\[\square\]

**Proof of Minkowski’s inequality for $p = 2$.**

$$\|X + Y\|_2^2 = E(X + Y)^2 = E(X^2) + E(Y^2) + 2E(XY) \leq \|X\|_2^2 + \|Y\|_2^2 + 2\|X\|_2\|Y\|_2 = (\|X\|_2 + \|Y\|_2)^2$$

\[\square\]
Sketch of proof for general \( p \geq 1 \). We will use the more general version of Jensen’s inequality: if \( \varphi : [0, \infty) \times [0, \infty) \rightarrow \mathbb{R} \) is convex then and \( X, Y \geq 0 \) then \( \varphi(E(X), E(Y)) \leq E(\varphi(X, Y)) \).

We apply this to the convex function
\[
\varphi(x, y) = -(x^{1/p} + y^{1/p})^p \quad x, y \geq 0
\]
We get
\[
E(\sqrt[p]{X} + \sqrt[p]{Y})^p \leq (\sqrt[p]{E(X)} + \sqrt[p]{E(X)})^p
\]
We now replace \( X, Y \geq 0 \) by \( |X|^p, |Y|^p \)

The following generalization of Cauchy-Schwarz inequality is often useful

**Theorem 5.16** (Hölder’s inequality). Suppose \( p, q > 1 \) are conjugate exponents \( 1/p + 1/q = 1 \). Then
\[
(5.16) \quad |E(XY)| \leq \|X\|_p \|Y\|_q
\]

**Sketch of proof.** We apply Jensen’s inequality to convex function \(-\sqrt[p]{x} \sqrt[q]{y}, x, y \geq 0 \). We get
\[
E\left(\sqrt[p]{X} \sqrt[q]{E(Y)}\right) \sqrt[p]{E(X)} \sqrt[q]{E(Y)}
\]
We then replace \( X, Y \geq 0 \) by \( |X|^p \) and \( |Y|^q \) to get \( E(XY) \leq E(|X||Y|) \leq \|X\|_p \|Y\|_q \).

**Other proofs.** The geometric mean is smaller than the arithmetic mean, so \( \alpha^{1/p} \beta^{1/q} \leq \alpha/p + \beta/q \).

Or, what is really the same, for \( 0 < u < 1 \) we have \( u^{1/q} \leq 1/p + u/q \) as by the mean value theorem applied to \( f(u) = u^{1/q} \) we have \( f(1) - f(u) = (1 - u)f'(\theta) = \frac{(1-u)\theta^{-1/p}}{q} \geq \frac{(1-u)}{q} \)

This gives \( |ab| \leq \left( \frac{|a|^p}{p} + \frac{|b|^q}{q} \right), \) and we can now modify proof \#3 of the Cauchy-Schwarz inequality.

**Another proof of Minkowski’s inequality for \( p > 1 \).** When \( p > 1 \) we have \( q = p/(p - 1) \). We apply monotonicity, linearity, and Hölder inequalities:
\[
E(|X+Y|^p) = E(|X+Y|^{p-1}|X+Y|) \leq E(|X+Y|^{p-1}|X|+|Y|) = E(\|X\||X+Y|^{p-1}) + E(\|Y\||X+Y|^{p-1}) \\
\leq (E(|X|^p)^{1/p}(E|X+Y|^{q(p-1)})^{1/q} + E(|Y|^p)^{1/p}(E|X+Y|^{q(p-1)})^{1/q} = (\|X\|_p + \|Y\|_p)E(|X+Y|^p)^{1/q}
\]
Here we used \( q(p-1) = p \) and \( 1 - 1/q = 1/p \).

**Another proof of Lyapunov’s inequality (5.13).** Write \( |X|^\alpha \) as product \( 1 \times |X|^\alpha \) and apply Hölder’s inequality with \( p = \beta/\alpha, q = 1 - 1/p = (\beta - \alpha)/\beta \). We get \( E|X|^\alpha \leq (E|X|^\beta)^{\alpha/\beta} \) which implies (5.13).

**Definition 5.4.** We say that random variables \( X_n \) converges to \( X \) converge in \( L_p \), if \( \|X_n - X\|_p \to 0 \). When \( p = 2 \) we also say that \( X_n \) converge in mean square.

It is clear that if \( X_n \to X \) and \( Y_n \to Y \) in \( L_p \) then \( X_n + Y_n \to X + Y \) in \( L_p \).
5. The law of large numbers

This is based on [Billingsley, Section 6]. Let $X_1, X_2, \ldots$ be a sequence of simple independent identically distributed random variables on some probability space $(\Omega, \mathcal{F}, P)$. Define $S_n = X_1 + \cdots + X_n$. Denote $m = E(X_n)$.

**Theorem 5.17.** $\frac{1}{n} S_n \to m$ with probability one.

**Convergence in probability.** In this proof we only show convergence in probability. Denote by $\sigma^2$ the variance of $X_1$. (Recall that simple random variables have all moments.) Then $\text{Var}(\frac{1}{n} S_n) = \frac{1}{n^2} \text{Var}(S_n) = \frac{\sigma^2}{n} \to 0$. Therefore, by Chebyshev’s inequality, for $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n} S_n - m\right| > \varepsilon\right) \leq \frac{\sigma^2}{n \varepsilon^2} \to 0$$

□

**Example 5.4** (The proof gives more!). We do not need independence: Suppose $\Omega = (0, 2\pi)$ with $P = \lambda/2\pi$. Let $X_k(\omega) = \cos(k\omega)$. Then $\frac{1}{n} S_n \to 0$ in probability.

**Example 5.5** (The proof gives more!). We do not need identical distributions: Suppose $X_n = a_n \gamma_n$ where $\gamma_n$ are independent mean zero variance 1. If $\{a_n\}$ is a bounded sequence then $\frac{1}{n} S_n \to 0$ in probability.

(Omitted in 2019) Proof moved to Chapter 8

**Convergence of a sub-sequence.** Here we show a specific subsequence, $\frac{1}{n} S_{n^2} \to m$ with probability one. Denote by $\sigma^2$ the variance of $X_1$. Then $\text{Var}(\frac{1}{n} S_n) = \frac{1}{n^2} \text{Var}(S_n) = \frac{\sigma^2}{n} \to 0$. Therefore, by Chebyshev’s inequality, for $\varepsilon > 0$,

$$\sum_n P\left(\left|\frac{1}{n^2} S_{n^2} - m\right| > \varepsilon\right) \leq \sum_n \frac{\sigma^2}{n^2 \varepsilon^2} < \infty.$$

Therefore, by Borel-Cantelli lemma, $P\left(\left|\frac{1}{n^2} S_{n^2} - m\right| > \varepsilon \text{ i.o.}\right) = 0$, i.e. $\frac{1}{n^2} S_{n^2} \to m$ with probability one. This in fact implies the result by the following reasoning, in which without loss of generality we assume $m = 0$. Given $n \in \mathbb{N}$ take $k = k(n)$ such that $k^2 \leq n < (k+1)^2$. Then, since $|X_j| \leq M$, for every $\omega \in \Omega$ we have

$$\frac{S_{k^2} - M(n - k^2)}{n} \leq \frac{S_n}{n} \leq \frac{S_{k^2} + M(n - k^2)}{n}$$

So

$$\frac{k^2 S_{k^2}}{n} - M(1 - \frac{k^2}{n}) \leq \frac{S_n}{n} \leq \frac{k^2 S_{k^2}}{n} + M(1 - \frac{k^2}{n})$$

Now $k = k(n) \to \infty$ so $k^2/(k+1)^2 \to 1$. Since $k^2 \leq n < (k+1)^2$ we see that $k^2/n \to 1$, and the result follows by squeezing principle, as $\frac{1}{n^2} S_{n^2} \to 0$ for all $\omega \in \Omega_0 \subset \Omega$ of probability one. □
Proof. Without loss of generality we can assume \( m = 0 \). (Replace \( X_n \) by \( X_n - m \).) We will use Borel-Cantelli lemma to verify that for every \( \varepsilon > 0 \), \( P\left( \frac{1}{n}|S_n| \geq \varepsilon \text{i.o.} \right) = 0 \). We use Markov’s inequality,

\[
P\left( \frac{1}{n}|S_n| \geq \varepsilon \right) \leq \frac{E[S_n^4]}{\varepsilon^4 n^4}
\]

We note that

\[
E[S_n^4] = \sum_{j_1,j_2,j_3,j_4=1}^{n} E[X_{j_1}X_{j_2}X_{j_3}X_{j_4}] = nE(X_1^4) + 3n(n-1)(E[X_1^2])^2 \leq Cn^2
\]

Thus \( \sum_n P\left( \frac{1}{n}|S_n| \geq \varepsilon \right) < \infty \). By Borel-Cantelli (Theorem 3.8) \( P\left( \frac{1}{n}|S_n| > \varepsilon \text{i.o.} \right) = 0 \), ending the proof. (See discussion of convergence with probability one in the proof of Proposition 4.14.) \( \square \)

6. Large deviations

Suppose \( X_n \) is Bin\((n,p)\) and let \( \hat{p}_n = X_n/n \) denote the sample proportion. Define

\[
I(x) = x \log x + (1-x) \log(1-x)
\]

Theorem 5.18. For \( x > p \),

\[
P(\hat{p}_n > x) \leq e^{-nI(x)}
\]

Proof. We apply Markov inequality to non-negative (simple) random variable \( \exp(tX_n) \), and choose optimal \( t \).

\[
P(\hat{p}_n > x) = P(X_n > nx) = P(\exp(tX_n) > \exp(nx))
\]

\[
\leq \frac{E[\exp(tX_n)]}{e^{nx}} = e^{-nx(t + \log(1 + p(e^t - 1)))}
\]

Let's now choose \( t \) that maximizes \( f(t) = (xt - \log(1 + p(e^t - 1))) \). Setting \( f'(t) = 0 \) we get \( t_0 = \log \frac{1-p}{1-x} \) and \( f(t_0) = I(x) \) as claimed.

The inequality is not true for \( x < p \), as then \( P(\hat{p}_n > x) \to 1 \) while the right hand side converges to 0.

\[\text{Can you spot the two places where the assumption } x > p \text{ is hiding in the proof?}\]

Remark 5.19. When \( x < p \) we need to give a bound for \( P(\hat{p}_n < x) \). (Better non-asymptotic estimates are known.)

Required Exercises

Exercise 5.1 (Statistics). Suppose that \( X \) is a simple random variable. Show that the number \( m = E(X) \) minimizes the function \( x \mapsto f(x) = E((X - x)^2) \).

Solution: Expanding the square, \( f(x) = EX^2 - 2xEX + x^2 = EX^2 - 2mx + x^2 \). So \( f'(x) = -2m + 2x = 0 \) and \( f'(x) = 0 \) for \( x = m \). Since \( f''(x) = 2 > 0 \), this is a minimum. 

\[\text{Quadratic Loss Function}\]
Exercise 5.2. Suppose that $X_n$ is a sequence of random variables such that $|X_n| \leq Y$ where $Y$ has Poisson distribution. Show that $\{X_n\}$ is stochastically bounded. (For definition, see Exercise 4.14.)

Solution:

\[
P(|X_n| > K) \leq P(Y > K)
\]

Since $F_Y(K) = P(Y \leq K) \to 1$ as $K \to \infty$, for every $\varepsilon > 0$ we can find $K$ such that $P(Y > K) < \varepsilon$. (The fact that $Y$ has Poisson distribution does not matter here!)

Exercise 5.3. Suppose $X_n$ takes values $\pm 1$ with probability $1-1/(2n)$ and value $n$ with probability $1/n$. Show that $X_n$ is stochastically bounded. (For definition, see Exercise 4.14.)

Solution:

\[
P(|X_n| > K) = \begin{cases} 
1/n & n > K \\
0 & n \leq K
\end{cases}
\]

So given $\varepsilon > 0$, choose $1/K < \varepsilon$. Then by the above formula, $P(|X_n| > K) = 0 < \varepsilon$ if $n < K$ and $P(|X_n| > K) = 1/n < 1/K < \varepsilon$ if $n > K$.

Exercise 5.4. Suppose that $X$ is a simple random variable which has non-negative integers $\{0, 1, 2, \ldots\}$ as values. Use the definition of $\mathbb{E}(X)$, not $(5.9)$, to prove that $\mathbb{E}(X) = \sum_{n=1}^\infty P(X \geq n)$.

(Of course, this is a finite sum!). Other similar exercises are possible, and using $(5.9)$ sometimes simplifies the solution, see Exercises 5.18, 5.20, 5.19 below.

Solution: Here is a (forbidden) solution that uses $(5.9)$: $EX = \int_0^\infty P(X > t)dt = \sum_{n=1}^\infty \int_{n-1}^n P(X > t)dt = \sum_{n=1}^\infty \int_{n-1}^n P(X \geq n)dt = \sum_{n=1}^\infty P(X \geq n)$

Exercise 5.5. Suppose $X$ is uniform $U(0, 1)$ random variable, $X_n$ is its approximation from the proof of Theorem 5.1. Compute $E(X_n)$ and its limit as $n \to \infty$.

Solution: $X_n = k - 1/2^n$ when $U \in [(k-1)/2^n, k/2^n)$ so $E(X_n) = \sum_{k=1}^{2^n} (k-1)/(2^n)^2$. It is simpler to denote $m = 2^n$ and re-write this as $\frac{1}{m^2} \sum_{k=1}^m (k-1) = \frac{m(m+1)}{2m^2} \to 1/2$ as expected.

Exercise 5.6. Suppose $X$ is Poisson random variable with parameter $\lambda$ (see Example 1.6 on page 7), written as $X = \sum_{k=0}^\infty kI_{A_k}$. Then as discrete monotone approximation we can take

\[
X_n = \sum_{k=0}^{n-1} kI_{A_k} + nI_{\bigcup_{k=n}^\infty A_k}.
\]

Compute $\lim_{n \to \infty} E(X_n)$.

Solution: Somewhat more generally, if $X = \sum_{k=0}^\infty kI_{A_k}$ then $X_n = \sum_{k=0}^{n-1} kI_{A_k} + nI_{\bigcup_{k=n}^\infty A_k}$ are simple random variables such that $X_n \uparrow X$. So $E(X) = \lim_{n \to \infty} \left( n \sum_{k=n}^\infty p_k + \sum_{k=1}^{n-1} kp_k \right)$

Exercise 5.7. For general $X \geq 0$, Remark 5.9 suggests a possible definition of the expected value via tail-integration formula.
(i) Suppose \( X \) is uniform \( U(0,1) \) random variable. Use this approach to see what this approach gives for \( \mathbb{E}[X] \).

(ii) Suppose \( X \) is exponential random variable, see Example 2.4 on page 18. Use this approach to see what this approach gives for \( \mathbb{E}[X] \).

(iii) Suppose \( X \) has CDF from Example 2.6 on page 19. Use this approach to see what this approach gives for \( \mathbb{E}[X] \).

Solution: Uniform: \( F(x) = 1 - x \) so \( \mathbb{E}(X) \) should be \( \int_0^1 (1 - t) \, dt = -(1 - t)^2/2 \big|_{t=0}^{t=1} = 1/2 \) as expected.

Exponential: With \( F(x) = 1 - e^{-\lambda x} \) the tail integration formula says that \( \mathbb{E}(X) \) should be \( \int_0^\infty e^{-\lambda t} \, dt = 1/\lambda \) as expected.

Exercise 5.8. Suppose \( 0 \leq X \leq 1 \) has cumulative distribution function \( F(x) \), and \( X_n \) is its approximation from the proof of Theorem 5.1. Express \( \mathbb{E}(X_n) \) solely in terms of \( F \).

Solution: 
\[
\mathbb{E}(X_n) = \sum_{k=1}^{n2^n} k - 1 - \frac{1}{2^n} \left( \lim_{x \uparrow \frac{k-1}{2^n}} F(x) - \lim_{x \uparrow \frac{k}{2^n}} F(x) \right)
\]

Exercise 5.9. Prove that for any simple r.v. \( X \) (positive or not) and any real numbers \( a, t \) we have

\[
P(X > t) \leq e^{-at} \mathbb{E} \exp(aX)
\]

Exercise 5.10. We say that random variables are centered if their mean is zero. We say that random variables \( X, Y \) are uncorrelated if \( \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \). Show that if \( X_1, X_2, \ldots \) are simple, (pairwise) uncorrelated, have the same variance \( \sigma^2 \), and centered then \( \frac{1}{n} S_n \to 0 \) in mean square.

Exercise 5.11. Show that \( L_p \)-convergence implies convergence in probability: if \( \|X_n - X\|_p \to 0 \) then \( X_n \xrightarrow{p} X \). (Thus together with Exercise 5.10 this establishes the so called weak law of large numbers: \( \frac{1}{n} S_n \xrightarrow{p} 0 \). Hint: the proof relies on a suitable application of (5.10).

Exercise 5.12. Suppose \( X_1, X_2, \ldots \) are independent uniformly bounded (say, \(|X_n| \leq 17\) for all \( n \)) mean zero (simple) random variables. Prove that

\[
\frac{1}{n^2} \sum_{j=1}^{n^2} X_j X_{j+1} \to 0
\]

with probability one.

Hint: Use Borel-Cantelli Lemma to verify that \( \Omega_0 = \{ \omega : \frac{1}{n^2} \sum_{j=1}^{n^2} X_j X_{j+1} \to 0 \} \) has probability one.
Exercise 5.13. Let $X, Y$ be simple random variables that (together) take values $0, 1, 2, \ldots, m$. Write

$$X = \sum_{j=0}^{m} jI_{A_j}, \quad Y = \sum_{j=0}^{m} jI_{B_j}.$$ 

Show that $\sigma(X, Y) = \sigma(A_0, A_1, \ldots, A_m, B_0, B_1, \ldots, B_m)$. Then describe $\sigma(Z)$ for $Z = X - Y$

Exercise 5.14 (Computer Science). Suppose $X_n$ is $Bin(n, 1/2)$. Apply (5.17) to sample proportion $X = X_n/n$ choosing $a$ in that will minimize the right hand side. State the resulting inequality in terms of a bound for $\frac{1}{n} \log P(\frac{1}{n}X_n < p)$, where $p < 1/2$. (Compare Theorem 5.18.)

Exercise 5.15. Is Theorem 5.7 true if we replace assumption $X_n \xrightarrow{P} X$ by a weaker condition $X_n \xrightarrow{P} X$?

Exercise 5.16. Is Theorem 5.7 true if we replace boundedness assumption by a weaker condition that $\{X_n\}$ is bounded in probability one?

Exercise 5.17. Is Theorem 5.7 true if we replace boundedness assumption by a weaker condition that $\{X_n\}$ is bounded in probability?

Exercise 5.18 (*). Suppose that $X$ is a simple random variable which has non-negative integers $\{0, 1, 2, \ldots \}$ as values. Use (5.9), or some other means, to prove that

$$\mathbb{E}(X^2) = \sum_{n=1}^{\infty} (2n - 1)P(X \geq n)$$

(Of course, this is a finite sum!).

\[ \text{Solution: Let's proceed with the right hand side: } \sum_{n=0}^{\infty} 2^n P(X \geq n + 1) = \sum_{n=0}^{\infty} 2^n \sum_{k=n+1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} P(X = k) \sum_{n=0}^{k-1} 2^n = \sum_{k=1}^{\infty} P(X = k)(2^k - 1) = \sum_{k=1}^{\infty} 2^k P(X = k) - \sum_{k=1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} 2^k P(X = k) - \sum_{k=0}^{\infty} 2^k (1 - p_0) = \sum_{k=0}^{\infty} 2^k P(X = k) - 1 = \mathbb{E}[X] - 1 \]

Exercise 5.19 (*). Suppose that $X$ is a simple random variable which has non-negative integers $\{0, 1, 2, \ldots \}$ as values. Prove that

$$\mathbb{E}[2^X] = 1 + \sum_{n=0}^{\infty} 2^n P(X \geq n + 1)$$

(Of course, this is a finite sum!).
Exercise 5.20 (**). Suppose that $X$ is a simple random variable which has positive integers \{1, 2, ...\} as values. Use (5.9), to prove that

$$
\mathbb{E}\left[\frac{1}{X}\right] = 1 - \sum_{n=1}^{\infty} \frac{1}{n(n+1)} P(X \geq n + 1)
$$

(Of course, this is a finite sum!).

Solution: Here one can proceed either with tail integration and change of variable as in Exercise 5.20, or work with the right hand side as in Exercise 5.19.

First method requires noticing that $P(1/X > t) = 0$ when $t > 1$, and then working out $\mathbb{E}[X] = \int_{0}^{\infty} P(1/X > t)dt = \int_{0}^{1} P(X < 1/t)dt = 1 - \int_{0}^{1} P(X > 1/t)dt = 1 + \int_{1}^{\infty} 1/u^2 P(X > u)du = 1 - \sum_{n=1}^{\infty} \int_{n}^{n+1} 1/u^2 P(X > u)du = 1 - \sum_{n=1}^{\infty} P(X \geq n + 1) \int_{n}^{n+1} 1/u^2du$...

The second method requires noticing the telescoping sum: $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} P(X \geq n + 1) = \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=n+1}^{\infty} P(X = k) = \sum_{k=2}^{\infty} P(X = k) \sum_{n=1}^{k-1} \frac{1}{k} = \sum_{k=2}^{\infty} P(X = k) \frac{1}{k} = 1 - \mathbb{E}[1/X]$.

Exercise 5.21. Complete details in the sketch of proof for Minkowski’s inequality.

Exercise 5.22. Complete details in the sketch of proof for H"older’s inequality.

Exercise 5.23. Show that $X_n \to X$ with probability 1 iff for every $\varepsilon > 0$ there exists $n$ such that $P(|X_k - X| < \varepsilon, n \leq k \leq m) \geq 1 - \varepsilon$ for all $m > n$.

Exercise 5.24 (*). Suppose $X_1, X_2, \ldots$ are independent uniformly bounded mean zero (simple) random variables. Prove that

$$
\frac{1}{n} \sum_{j=1}^{n} X_j X_{j+1} \to 0
$$

with probability 1.

Hint: Prove (5.18) first. Noting that every $n$ can be put into an interval $k^2 \leq n < (k+1)^2$, use the uniform bound on $X_j$ to show that if (5.18) hold for some $\omega \in \Omega$, this implies convergence in (5.19) for the same $\omega$.

Exercise 5.25. Suppose $X$ has mean $m$ and variance $\sigma^2$. For $\alpha \geq 0$, prove Cantelli’s inequality

$$
P(X - m \geq \alpha) \leq \frac{\sigma^2}{\sigma^2 + \alpha^2}
$$

Deduce that

$$
P(|X - m| \geq \alpha) \leq \frac{2\sigma^2}{\sigma^2 + \alpha^2}
$$

When is this better than Chebyshev’s inequality?

Hint: Assume $m = 0$. $P(X \geq \alpha) \leq P((X + x)^2 \geq (\alpha + x)^2)$. Apply Markov’s inequality, minimize over $x > 0$. 88
1. Approximation by simple random variables

**Theorem 6.1** (Measurability Theorem). Suppose $X(\omega) \geq 0$ for all $\omega$. Then $X$ is measurable iff there exist simple random variables $X_n$ such that

$$0 \leq X_n(\omega) \leq X_{n+1}(\omega) \quad \text{and} \quad X_n(\omega) \uparrow X(\omega)$$

**Proof.** By Remark 4.17, the limit $X$ of simple random variables is measurable. Conversely, if $X$ is measurable, define

$$X_n(\omega) = \begin{cases} 0 & \text{if } X < \frac{1}{2^n} \\ \frac{1}{2^n} & \text{if } \frac{1}{2^n} \leq X < \frac{2}{2^n} \\ \vdots & \\ \frac{n-1}{2^n} & \text{if } \frac{n-1}{2^n} \leq X < \frac{n}{2^n} \\ n & \text{if } X \geq n2^n \end{cases}$$

Notice that $X_n$ is measurable, as it is a composition of measurable functions, $X_n(\omega) = \psi_n(X(\omega))$, with

$$\psi_n(x) = n2^n I_{[n2^n, \infty)}(x) + \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{[\frac{k-1}{2^n}, \frac{k}{2^n})}(x).$$
The following is a generalization of Theorem 5.1.

**Theorem 6.2.** Let \( X = (X_1, \ldots, X_k) \) be a random vector. Then the following are equivalent

(i) random variable \( Y \) is measurable with respect to \( \sigma(X) = \sigma(X_1, \ldots, X_k) \)

(ii) there exists a measurable map \( f : \mathbb{R}^k \to \mathbb{R} \) such that \( Y = f(X_1, \ldots, X_k) \)

Proof. Noting that \( A \in \sigma(X_1, \ldots, X_k) \) if and only if \( A = X^{-1}(U) \) for some Borel subset \( U \) of \( \mathbb{R}^k \), it is clear that \( Y^{-1}(V) = X^{-1} \circ f^{-1}(V) \in \sigma(X_1, \ldots, X_k) \).

Conversely, if \( Y \) is a simple random variable with distinct values \( y_1, \ldots, y_m \) then the sets \( A_j = \{ \omega : Y(\omega) = y_j \} \in \sigma(X_1, \ldots, X_k) \) are disjoint so \( A_j = \{ \omega : X(\omega) \in U_j \} \) with Borel sets \( U_j \subset \mathbb{R}^k \). In this case, \( Y = f(X) \) where \( f = \sum y_j I_{U_j} \), as \( X \in U_j \cap U_j \) cannot occur.

In general, take simple random variables such that \( Y_n(\omega) \to Y(\omega) \). By the construction, \( Y_n \) are \( \sigma(X) \)-measurable so \( Y_n = f_n(X) \).

Consider \( M \subset \mathbb{R}^k \) such that \( f_n(x) \) converge for \( x \in M \). Then \( M \) is measurable, and

\[
f(x) = \begin{cases} 
\lim_{n \to \infty} f_n(x) & \text{if } x \in M \\
0 & \text{otherwise}
\end{cases}
\]

□

2. Expected values

Recall that the expected value of a simple random variable \( X = \sum x_j I_{A_j} \)

\[
\mathbb{E}[X] = \int X dP = \int_{\Omega} X(\omega) P(d\omega)
\]

is defined as \( \sum x_j P(A_j) \). We now extend this definition to non-negative random variables:
**Definition 6.1.** If \( X \geq 0 \) we define \( \mathbb{E}[X] \in [0, \infty] \) as \( \lim_{n \to \infty} \mathbb{E}[X_n] \), where \( X_n \uparrow X \) are non-negative simple random variables.

**Proposition 6.3.** If \( X \geq 0 \) then \( \mathbb{E}[X] \) is well defined.

**Proof.** We need to show that if \( X_n \uparrow X \) and \( Y_m \uparrow X \) then the limits coincide.

To do that, we note that for each \( n \) we have
\[
 Z_m := X_n \wedge Y_m \uparrow X_n \wedge X = X_n \quad \text{as} \quad m \to \infty
\]
Since \( Z_m \leq Y_m \), therefore by monotonicity (Proposition 5.6) the limits (possibly infinite) exist and we have \( \lim_{m \to \infty} \mathbb{E}[Y_m] \geq \lim_{m \to \infty} \mathbb{E}(Z_m) \). Since \( Z_m \leq X_n \leq K \) for all \( m \), by Theorem 5.7
\[
 \lim_{m \to \infty} \mathbb{E}(Z_m) = \mathbb{E}(X_n)
\]
Thus \( \lim_{n \to \infty} \mathbb{E}(X_n) \leq \lim_{m \to \infty} \mathbb{E}(Y_m) \). By symmetry, equality follows. \( \square \)

The general definition uses the decomposition \( X = X^+ - X^- \) into the positive and negative parts.

**Definition 6.2.** We say that \( X \) is an integrable random variable if \( \mathbb{E}(X^+) < \infty \) and \( \mathbb{E}(X^-) < \infty \). For integrable random variables, we define \( \mathbb{E}[X] = \mathbb{E}(X^+) - \mathbb{E}(X^-) \).

We note that \( X \) is integrable means that \( \mathbb{E}(|X|) < \infty \).

We also write \( \int_{\Omega} XdP \) or \( \int X(\omega)P(d\omega) \) instead of \( \mathbb{E}[X] \). The integral notation is sometimes more convenient when integrating over more complicated sets, like in
\[
 \int_A X(\omega)P(d\omega) := \mathbb{E}[XI_A]
\]
or
\[
 \mathbb{E}[I_{|X|>1}X] = \int_{|X|>1} X(\omega)P(d\omega)
\]
or when we want to point out dependence on the measure, as in change of variable formula (6.5).

We note that if \( P(X = Y) = 1 \) and one of them is integrable then the second one is integrable and \( \mathbb{E}[X] = \mathbb{E}[Y] \). Indeed, if \( P(A) = 1 \) then \( \mathbb{E}[XI_A] = \mathbb{E}[X] \) as this holds for simple random variables.

**2.1. Expected values and limits.** This section is based on [Billingsley, Section 16].

**Theorem 6.4** (Monotone Convergence Theorem). If \( X_n \geq 0 \) and \( X_n \uparrow X \) (with probability one) then \( \mathbb{E}[X_n] \to \mathbb{E}[X] \). (Here we allow infinite values.)

First we prove a lemma:

**Lemma 6.5.** If \( 0 \leq X \leq Y \), then \( \mathbb{E}[X] \leq \mathbb{E}[Y] \).

**Proof.** Indeed, if \( X_n \uparrow X \) and \( Y_n \uparrow Y \) are simple then \( Z_n = \min\{X_n, Y_n\} \uparrow X \) and \( \mathbb{E}(Z_n) \leq \mathbb{E}(Y_n) \). \( \square \)

**Proof of Theorem 6.4.** To prove the theorem, we construct a sequence of simple random variables \( Z_m \) such that \( Z_m \uparrow X \). The starting point of the construction is the array of simple random
variables $Y_{m,n} \uparrow X_n$ as $m \to \infty$.

$$\begin{array}{cccc}
Y_{1,1} & Y_{2,1} & \cdots & Y_{m,1} & \cdots \nearrow X_1 \\
Y_{1,2} & Y_{2,2} & \cdots & Y_{m,2} & \cdots \nearrow X_2 \\
Y_{1,3} & Y_{2,3} & \cdots & Y_{m,3} & \cdots \nearrow X_3 \\
\vdots & \vdots & \cdots & \vdots & \\
Y_{1,m} & Y_{2,m} & \cdots & Y_{m,m} & \cdots \nearrow X_m \\
Y_{1,m+1} & Y_{2,m+1} & \cdots & Y_{m,m+1} & \cdots \nearrow X_{m+1} \\
\vdots & \vdots & \cdots & \vdots & \\
Y_{1,n} & Y_{2,n} & \cdots & Y_{m,n} & \cdots \nearrow X_n \\
\vdots & \vdots & \cdots & \vdots & \\
Z_1 & Z_2 & \cdots & Z_m & \nearrow X^* \end{array}$$

We take $Z_m = \max_{k \leq m} Y_{m,k}$, so $Z_1 = Y_{1,1}$, $Z_2 = \max \{Y_{2,1}, Y_{2,2}\}$, $Z_3 = \max \{Y_{3,1}, Y_{3,2}, Y_{3,3}\}$, etc.

Note that $Z_2 \geq Y_{2,1} \geq Y_{1,1} = Z_1$. Similarly, $Z_3 \geq Y_{3,1} \geq Y_{2,1}$ and $Z_3 \geq Y_{3,2} \geq Y_{2,2}$, so $Z_3 \geq \max \{Y_{2,1}, Y_{2,2}\} = Z_2$.

In general, $Z_m \leq Z_{m+1}$, so $Z_m \nearrow X^*$ for some $X^*$ (that possibly could be infinite).

It is clear that $X_1 = \lim_{m \to \infty} Y_{m,1} \leq \lim_{m \to \infty} Z_m = X^*$. Similarly, $X_2 = \lim_{m \to \infty} Y_{m,2} \leq \lim_{m \to \infty} Z_m = X^*$. In general, $X_n \leq X^*$. So $X = \lim_{n \to \infty} X_n \leq X^*$.

Next we note that $Z_1 = Y_{1,1} \leq X_1$, similarly, $Z_2 = \max \{Y_{2,1}, Y_{2,2}\} \leq \max \{X_1, X_2\} = X_2$. In general, $Z_m \leq X_m$, so taking the limits, we get $X^* \leq X$. (In particular, there is no danger of $X^*$ being infinite.)

Combining both inequalities, we see that $X = X^*$ so $\mathbb{E}[X] = \lim_{n \to \infty} \mathbb{E}(Z_n)$. Furthermore, $X_n \leq X$ so by Lemma $\mathbb{E}(X_n) \leq \mathbb{E}[X]$ and $\lim_{n \to \infty} \mathbb{E}(X_n) \leq \mathbb{E}[X]$. On the other hand, $Z_m \leq X_m$ implies that $\lim_{n \to \infty} \mathbb{E}(X_m) \leq \lim_{m \to \infty} \mathbb{E}[Z_m]$.

It is clear that $Z_m \uparrow X^*$ are simple random variables, so $\mathbb{E}[X] = \mathbb{E}(X^*) = \lim_{n \to \infty} \mathbb{E}(Z_n)$.

**Corollary 6.6.** If $\sup_n E(|X|I_{|X|<n}) < \infty$ then $X$ is integrable.

**Proof.** $|X|I_{|X|<n} \nearrow |X|$. $\square$

**Remark 6.7.** With appropriate restrictions on the class of simple functions, Definition 6.1 can also be used to define $\int_{\mathbb{R}} f(x)dx$ for positive $f$. Monotone convergence theorem then holds true, too.

Sometimes it is convenient to prove integrability using non-monotone sequences.

**Theorem 6.8** (Fatou’s lemma). For non-negative $X_n$ we have

$$E \left( \liminf_{n \to \infty} X_n \right) \leq \liminf_{n \to \infty} E(X_n)$$

**Proof.** If $Y_n = \inf_{k \geq n} X_k$ then $0 \leq Y_n \uparrow \liminf_{n \to \infty} X_n$. So by Theorem 6.4 we have $E \left( \liminf_{n \to \infty} X_n \right) = \lim_{n \to \infty} EY_n$. Since $X_n \geq Y_n$, we have $E(X_n) \geq E(Y_n)$, so $\liminf_{n \to \infty} E(X_n) \geq \lim_{n \to \infty} EY_n = E \left( \liminf_{n \to \infty} X_n \right)$.

**Theorem 6.9** (Lebesgue’s dominated convergence theorem). If $X_n \overset{P}{\to} X$ and there exists an integrable random variable $Z$ such that $|X_n| \leq Z$, then $E(X_n) \to E[X]$. 

Lemma 6.10. If $X$ is integrable then $|EX| \leq E|X|$

**Proof.** $EX = EX^+ - EX^- \leq EX^+ \leq EX^+ + EX^- = E|X|$ and $EX = EX^+ - EX^- \geq -EX^+ - EX^- = -E|X|$. □

Lemma 6.11. If $X, Y \geq 0$ then $E(X + Y) = EX + EY$

**Proof.** Take simple $X_n \uparrow X$ and $Y_n \uparrow Y$. Then $X_n + Y_n \uparrow X + Y$ and using (5.5) we have $E(X + Y) = \lim_{n \to \infty} E(X_n + Y_n) = \lim_{n \to \infty} EX_n + \lim_{n \to \infty} EY_n = EX + EY$. □

Lemma 6.12. If $X$ is integrable random variable then for every $\varepsilon > 0$ there exists $K$ such that $\int_{|X| > K} |X|dP < \varepsilon$.

**Proof.** We note that $|X|I_{|X| < n} \uparrow |X|$ and $|X| = |X|I_{|X| < n} + |X|I_{|X| \geq n}$.

Therefore $E(|X|) = E(|X|I_{|X| < n}) + E(|X|I_{|X| \geq n})$ and hence

$$E(|X|I_{|X| \geq n}) = E(|X|) - E(|X|I_{|X| < n}) \to 0$$

by monotone convergence theorem. □

**Proof of Theorem 5.9.** This proof is based on (improved) estimates that we already used in the proof of Theorem 5.7. First we note$^1$ that $|X_n - X| \leq |X_n| + |X| \leq 2Z$.

Given $\varepsilon > 0$ choose $K$ such that $\int_{2Z > K} ZdP < \varepsilon/2$, see Corollary 6.12. Then by Lemma 6.11 we have $E(|X_n - X|) = E(|X_n - X|I_{|X_n - X| \leq K}) + E(|X_n - X|I_{|X_n - X| > K})$. From Lemma 6.10 we get $E(|X_n - X|I_{|X_n - X| > K}) \leq 2E(|Z|I_{|X_n - X| > K}) \leq 2E(|Z|I_{|X_n| + |X| > K}) \leq 2E(ZI_{2Z > K}) < \varepsilon$. So

$$E(|X_n - X|) \leq E(|X_n - X|I_{|X_n - X| \leq K}) + \varepsilon$$

Now, again by Lemma 6.11

$$E(|X_n - X|I_{|X_n - X| \leq K}) = E(|X_n - X|I_{|X_n - X| \leq K, |X_n - X| > \varepsilon}) + E(|X_n - X|I_{|X_n - X| \leq K, |X_n - x| \leq \varepsilon}) \leq KP(|X_n - X| > \varepsilon) + \varepsilon.$$

Thus $E(|X_n - X|) \leq K(P(|X_n - X| > \varepsilon) + 2\varepsilon \to 2\varepsilon$ as $n \to \infty$. Since $\varepsilon > 0$ is arbitrary this shows that $\lim_{n \to \infty} E(|X_n - X|) = 0$. □

$^1$By Proposition 4.15, $|X| = \lim_{k \to \infty} |X_{n_k}| \leq Z$
Remark 6.13. Note that the assumptions in Theorem 6.9 are weaker than "the standard formulation" and can be further weakened. Firstly, the convergence is in probability. Secondly, the proof uses random variable $Z$ only to deduce that for every $\varepsilon > 0$ there exists $K$ such that $\int_{|X_n|>K} |X_n|P(d\omega) < \varepsilon$. This property is called uniform integrability.

Definition 6.3. A family $\{X_n\}$ of random variables is uniformly integrable if

$$\lim_{k \to \infty} \sup_n \int_{|X_n|>K} |X_n|dP = 0$$

Theorem 6.14 (DCT++). If $X_n \overset{P}{\to} X$ and $\{X_n\}$ is uniformly integrable, then $E(X_n) \to E(X)$.

Uniform integrability can be verified from higher moments.

Proposition 6.15. Suppose that $\sup_n E(X^2_n) < \infty$. Then $\{X_n\}$ is uniformly integrable.

Proof. Denote $M = \sup_n E(X^2_n) < \infty$. Given $\varepsilon > 0$, choose $K$ such that $M/K < \varepsilon$. By Cauchy-Schwarz and Markov inequalities

$$\int_{|X_n|>K} |X_n|dP \leq (E(I_{|X_n|>K})E(X^2_n))^{1/2} \leq \sqrt{M \sqrt{P(|X_n|>K)}} \leq \sqrt{M \sqrt{E[|X_n|^2]/K^2}} \leq M/K < \varepsilon$$

Corollary 6.16. If $X_n \overset{P}{\to} X$ and $\sup_n E|X_n|^2 < \infty$ then $E(X_n) \to EX$.

Proof. Step 1: integrability. We first need to verify that $EX^2 < \infty$. Choose a subsequence $X_{n_k} \to X$ By Fatou’s lemma, $EX^2 = E \lim \inf_{n \to \infty} X^2_{n_k} \leq \lim \inf_{n \to \infty} EX^2_{n_k} \leq \sup_n EX^2_n < \infty$.

Step 2: reduction. By Minkowski’s inequality, $\sup_n \|X_n - X\|_2 \leq \sup_n \|X_n\|_2 + \|X\|_2 < \infty$, so without loss of generality we can work with $\tilde{X}_n := X_n - X \to 0$ and it is enough to show that $E|\tilde{X}_n| \to 0$.

Step 3: repeat DCT proof. Next, we repeat the proof of Lebesgue dominated convergence theorem, with a slight modification: Let write $X_n$ for $\tilde{X}_n$. By Cauchy-Schwarz, and Markov inequalities $E|X_n|I_{|X_n|>K} \leq (E|X^2_n|)^{1/2} \sqrt{P(|X_n|>K)} \leq \frac{EX^2_n}{K} \leq M/K$. So for every $\varepsilon > 0$ we can find $K = M/\varepsilon$ such that $E|X_n|I_{|X_n|>K}$. Therefore, informally (as we do not yet have linearity!)

$$E|X_n| = E|X_n|I_{|X_n|>\varepsilon} + E|X_n|I_{|X_n|\leq \varepsilon} \leq K \sqrt{P(|X - n| > \varepsilon)} + \varepsilon$$

This shows that $\limsup_{n \to \infty} E|X_n| \leq \varepsilon$ for every $\varepsilon > 0$.

Example 6.1. Suppose $X_n = n^2I_{\{0,1/n\}}$ on $\Omega = [0,1]$ with Lebesgue measure. Then $X_n \to 0$ as $n \to \infty$ with probability one (in fact, for all $\omega \in \Omega$ except one point), but $E(X_n) = n \to \infty$.

Suppose $X_n = nI_{\{0,1/n\}}$ on $\Omega = [0,1]$ with Lebesgue measure. Then $X_n \to 0$ as $n \to \infty$ with probability one (in fact, for all $\omega \in \Omega$ except one point), but $E(X_n) = 1$ does not converge to 0.

Theorem 6.17. Expected value has the following properties:

(i) If $X \geq 0$ then $0 \leq E[X] \leq \infty$

(ii) If $X$ is integrable, then $E|X| \leq E(|X|)$

(iii) If $X,Y$ are integrable, then $X + Y$ is integrable and $E(X + Y) = E[X] + E(Y)$

(iv) If $X \leq Y$ are integrable then $E[X] \leq E(Y)$
Proof. Proof of (iii): $|X + Y| \leq |X| + |Y|$ so integrability follows. Now write $X = X_+ - X_-$
$Y = Y_+ - Y_-$ and take $X_n^+ - X_n^- \to X$ and $Y_n^+ - Y_n^- \to Y$ to be the simple function approximations
to $X^+, X^-$ and $Y^+, Y^-$. Then $|X_n + Y_n| \leq |X| + |Y|$ so by Lebesgue dominated convergence theorem
(Theorem 6.9) we can pass to the limit as $n \to \infty$ in

$$E(X_n) + E(Y_n) = E(X_n + Y_n)$$

\[ \square \]

2.2. Discrete random variables. Recall that a simple random variable has a finite number of values. A discrete random variable has a countable number of values.

Proposition 6.18. Suppose $X$ is a discrete random variable with $p_k = P(X = x_k)$ such that
$\sum p_k = 1$. If the series $\sum |x_k|p_k$ converges then $X$ is integrable and

$$E[X] = \sum_k x_k p_k$$

Proof. To see how this holds, we first consider $X \geq 0$ and write $X = \sum_{k=1}^{\infty} x_k I_{A_k}$ with $A_k = \{\omega : X(\omega) = x_k\}$. Then the simple random variable approximation is $X_n = \sum_{k=1}^{n} x_k I_{A_k} \uparrow X$ and $E[X] = \lim_{n \to \infty} E(X_n) = \lim_{n \to \infty} \sum_{k=1}^{n} x_k p_k = \sum_{k=1}^{\infty} x_k p_k$.

This shows that $E(|X|) < \infty$ and then $E(X^+ - X^-) = \sum_{k:x_k>0} x_k p_k - \sum_{k:x_k<0} (-x_k) p_k = \sum_{k=1}^{\infty} x_k p_k$ as the terms of the absolutely convergent series can be rearranged in any order. \[ \square \]

Example 6.2. Binomial distribution: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $k = 0, 1, \ldots, n$ (see Example 1.5) has mean $\mathbb{E}[X] = np$ and variance $\sigma^2 = np(1-p)$.

Example 6.3. Geometric distribution: $P(X = k) = p(1 - p)^{k-1}$, $k = 1, 2, \ldots$ is a special case of the negative binomial distribution from Example 1.7. The mean is $\mathbb{E}[X] = \frac{1}{p}$ and the variance is $\sigma^2 = \frac{1-p}{p^2}$.

Example 6.4. Poisson distribution: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, see Example 1.6. The mean is $\mathbb{E}[X] = \lambda$ and the variance is $\sigma^2 = \lambda$.

(6.4)

(6.5)

2.3. Change of measure and densities.

Proposition 6.19. If $X : \Omega \to \mathbb{R}$ is integrable and has cumulative distribution function $F$, then
$\mathbb{E}[X] = \int_{\mathbb{R}} x dF(x)$, where $dF$ is a probability measure on $\mathbb{R}, \mathcal{B}$ induced by $X$. More generally, if $g : \mathbb{R} \to \mathbb{R}$ then $E(g(X)) = \int_{\mathbb{R}} g(x) dF(x)$.
Proof. Here $\Omega' = \mathbb{R}$ with $P' = dF$, $T(\omega) = X(\omega)$ and $X'(\omega') = \omega'$.

Sketch of direct proof. Here we prove the case of general $g$ and then deduce the case $g(x) = x$.

$$\int_{\Omega} (g(X(\omega))) dP = \int_{\mathbb{R}} g(x) dF(x)$$

is first checked for simple $g$, then for positive $g$, then for general $g$.

The multivariate version of Corollary 6.19 is as follows.

**Corollary 6.20.** Suppose $(X_1, \ldots, X_k)$ has distribution $\mu$ and $g : \mathbb{R}^k \to \mathbb{R}$ is measurable. Then

$$E(g(X_1, \ldots, X_k)) = \int_{\mathbb{R}^k} g(x) \mu(dx)$$

provided the integrals exist.

2.3.1. **Densities.** Of special interest are cumulative distribution functions such that $F(x) = \int_{-\infty}^x f(y) dy$ where $f(y) \geq 0$ is called the density function, i.e. a non-negative measurable and integrable function that integrates to 1. (We do not assume continuity! Improper Riemann integrals are OK here.)

**Proposition 6.21.** If random variable $X$ is integrable and has cumulative distribution function $F(x) = \int_{-\infty}^x f(y) dy$ then

$$(6.6) \quad E[X] = \int_{\mathbb{R}} x f(x) dx.$$  

To prove this, we consider separately $X^+$ and $X^-$.  

We decompose $X = X^+ - X^-$ and approximate $\psi_n(X^+) \uparrow X^+$ with $\psi_n(x) = (k - 1)/2^n$ on $((k - 1)/2^n, k/2^n]$, $\psi_n(x) = 0$ for $x < 0$, compare (6.1). Then

$$E(\psi_n(X^+)) = \sum_{k=1}^{2^n} \frac{k-1}{2^n} \left( F\left(\frac{k}{2^n}\right) - F\left(\frac{k-1}{2^n}\right) \right) = \int_{\mathbb{R}} \psi_n(x) f(x) dx.$$  

Taking the limit, by monotone convergence theorem (see Remark 6.7) we get $E(X^+) = \int_{\mathbb{R}} x^+ f(x) dx$ and hence $E(X) = \int_{\mathbb{R}} (x^+ - x^-) f(x) dx = \int_{\mathbb{R}} x f(x) dx$

**Example 6.5.** Uniform density $U[a, b]$ is $f(x) = \frac{1}{b-a} I_{[a, b]}(x)$. The mean and the variance are $m = (a + b)/2$, $\sigma^2 = (b - 1)^2/12$.

**Example 6.6.** Exponential distribution: $F(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$. The density is $f(x) = e^{-x} I_{[0, \infty)}(x)$. The mean and the variance are $m = 1$, $\sigma^2 = 1$.

**Example 6.7.** Standard normal density: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The mean and the variance are $m = 0$ and $\sigma^2 = 1$. 

2.3.2. Multivariate densities. Similar approximation argument shows that if $\mu(dx) = f(x)dx$ has the density with respect to Lebesgue measure on $\mathbb{R}^k$ then
\[ \mathbb{E}[g(X_1, \ldots, X_k)] = \int_{\mathbb{R}^k} g(x) f(x) dx \]
In particular, $\text{cov}(X_1, X_2) = \int_{\mathbb{R}^2} (x - m_1)(y - m_2) f(x_1, x_2) dx_1 dx_2$

**Example 6.8.** Uniform distribution on a disk: $f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & x^2 + y^2 > 1 \end{cases}$

3. Inequalities

This is based on [Billingsley, Section 21].

Identities and inequalities from previous chapter extend to general random variables.

(Omitted in 2019)

**Theorem 6.22** (Tail integration formula). If $X \geq 0$ then $\mathbb{E}[X] = \int_0^\infty P(X > t) dt$

**Informal proof.** The formula holds true for $X_n \uparrow X$ simple. Noting that $I_{X_n > t} \nearrow I_{X > t}$ we get $P(X_n > t) \uparrow P(X > t)$, the result follows from the monotone convergence theorem applied to $f_n(t) = P(X_n > t)$, see Remark 6.7. This is an informal proof because it uses monotone convergence theorem for a Lebesgue measure on $[0, \infty)$ \hfill \square

Markov’s and Chebyshev’s inequalities (5.10) and (5.11) extend to all random variables.

**Proposition 6.23.** For $X \geq 0$ and $\alpha > 0$
\[ P(X \geq \alpha) \leq \frac{1}{\alpha} \mathbb{E}[X] \]
For $\alpha > 0$, and square-integrable $X$ with mean $m = \mathbb{E}[X]$
\[ P(|X - m| \geq \alpha) \leq \frac{\text{Var}(X)}{\alpha^2}. \]

**Proof.** For $X \geq 0$ and $\alpha > 0$ we have
\[ X \geq \alpha I_{X \geq \alpha} \geq \alpha I_{X \geq \alpha} \]
Integrating both sides with respect to $dP$ we get (5.10). \hfill \square

**Theorem 6.24** (Jensen’s inequality). If $\varphi$ is convex and both $X$ and $\varphi(X)$ are integrable then $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.

**Proof.** Here we use the following property of convex functions: $\varphi(x)$ is the supremum of linear functions; in fact, for every point $(x_0, \varphi(x_0))$ on the graph of $\varphi$ there is a line $ax + b$ that does through this point and lies below the graph of $\phi$.

We apply this with $x_0 = \mathbb{E}[X]$. Then $ax + b \leq \varphi(x)$ so $E(aX + b) \leq E(\varphi(X))$. But $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b = ax_0 + b = \varphi(x_0) = \varphi(\mathbb{E}[X])$.\hfill \square
6. Integration

Figure 2. Graph of $\varphi(x)$ with supporting line $y = ax + b$ at point $(x_0, \varphi(x_0))$. With $x_0 = EX$ we have $\varphi(x_0) = ax_0 + b = E[aX + b] \leq E\varphi(X)$.

Next we confirm that Cauchy-Schwarz inequality (5.14) holds for all square-integrable $X, Y$.

**Theorem 6.25** (Cauchy-Schwarz). If $X, Y$ are square integrable, then $XY$ is integrable and

$$|E(XY)| \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}$$

**Proof.** Since $2|XY| \leq |X|^2 + |Y|^2$ we see that $XY$ is integrable. Once we know integrability, we can use approximation by simple random variables $X_n \uparrow |X|$ and $Y_n \uparrow |Y|$ so that $X_nY_n \uparrow |X||Y|$. Then

$$|E(XY)| \leq E(|X||Y|) = \lim_{n \to \infty} E(X_nY_n) \leq \lim_{n \to \infty} (E(X_n^2)^{1/2}E(Y_n^2)^{1/2}.$$

Alternatively, $(X + tY)^2$ is integrable as a sum of three integrable functions (Theorem 6.17(iii)) so the previous proof based on positivity of polynomial $p(t) = E(X^2+Y^2)$ applies.

Similarly, Hölder’s inequality (5.16) holds, and can be deduced either from approximation by simple random variables or from numerical inequality that $|XY| \leq |X|^p/p + |Y|^q/q$ if $1/p + 1/q = 1$. So $XY$ is integrable and the previous proof can be repeated.

We remark that if $E|X|^p < \infty$ for some $p > 1$ then $E|X| < \infty$.

3.1. $L_p$-norms. As before, for $p \geq 1$ we define the $L_p$-norm of $X$ as

$$\|X\|_p = \sqrt[p]{E(|X|^p)}$$

Lyapunov’s inequality says that of $p_1 \leq p_2$ then $\|X\|_{p_1} \leq \|X\|_{p_2}$. This includes the statement that if $E|X|^p < \infty$ for some $p > 1$ then $X$ is integrable! In particular, $\|X\|_1 \leq \|X\|_2$.

Hölder’s inequality says that if $\|X\|_p < \infty$ and $\|Y\|_q < \infty$ for $1/p + 1/q = 1$ with $p > 1$ then $XY$ is integrable and

$$E(XY) \leq \|X\|_p\|Y\|_q$$
4. Independent random variables

If \( \|X\|_p < \infty \) and \( \|Y\|_p < \infty \) then \textit{Minkowski’s inequality} says that
\[
\|X + Y\|_p \leq \|X\|_p + \|Y\|_p
\]
This can be deduced from (5.15) for simple random variables by invoking the dominated convergence theorem, using the bound
\[
2 \|X + Y\|_p \leq 2^p (\|X\|^p + \|Y\|^p)
\]

There is also an \( L_\infty \) norm, defined as an essential supremum:
\[
\|X\|_\infty = \inf{\{k : P(\|X\| > k) = 0}\}
\]

Then Hölder’s inequality extends to \( p = 1 \):
\[
|E(XY)| \leq \|X\|_1 \|Y\|_\infty
\]

**Definition 6.4.** For \( p \geq 1 \) we say that \( X_n \to X \) in \( L_p \) if \( \|X_n - X\|_p \to 0 \).

\( L_1 \)-convergence and \( L_2 \)-convergence are often used. The \( L_2 \)-convergence is also called mean square convergence.

**Proposition 6.26.** If \( X_n \to X \) in \( L_p \) then \( X_n \xrightarrow{P} X \).

**Proof.** This is just an exercise on the use of (5.10).

4. Independent random variables

**Proposition 6.27.** Suppose \( X_1, \ldots, X_n \) are independent and integrable. Then \( Z = X_1X_2 \ldots X_n \) is integrable and
\[
E(Z) = E(X_1)E(X_2) \ldots E(X_n)
\]

This follows from the fact (5.8) for simple random variables. For simplicity of notation consider a pair of independent \( X = X^+ - X^-, Y = Y^+ - Y^- \). Then simple random variables \( X_n = \psi_n(X^+) - \psi_n(X^-) \) and \( Y_n = \psi_n(Y^+) - \psi_n(Y^-) \) are independent, \( |X_n| \uparrow |X| \) and \( |Y_n| \uparrow |Y| \) so \( |X_nY_n| \uparrow |XY| \) and \( E(|X||Y|) = EE(|X|)|Y| < \infty \).

Then \( X_nY_n \to XY \) and \( |X_nY_n| \leq |XY| \) so by dominated convergence theorem \( E(X_nY_n) \to E(XY) \). But for simple random variables \( E(X_nY_n) = E(X_n)E(Y_n) \to E[X]E(Y) \), again by dominated convergence theorem.

**4.1. Variances, Covariances.** For square integrable random variables, we define \( \text{Var}(X) \) and \( \text{Cov}(X,Y) \). For independent \( X_1, \ldots, X_n \) we have
\[
\text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)
\]

Somewhat more generally,
\[
\text{Var}\left( \sum_{j=1}^n X_j \right) = \sum_{i=1}^n \text{Var}(X_j) + 2 \sum_{i<j} \text{Cov}(X_i, X_j)
\]

**Proof.** \( |X + Y|^p \leq (|X| + |Y|)^p \leq (2|X| \vee |Y|)^p = 2^p|X|^p \vee |Y|^p \leq 2^p(|X|^p + |Y|^p) \) \( \square \)
For square-integrable random variables $X, Y$, the correlation coefficient is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Cauchy-Schwarz inequality implies that $|\rho| \leq 1$.

5. Moment generating functions

**Definition 6.5.** The moment generating function is

$$M(s) = \mathbb{E}(e^{sx})$$

for all $s \in \mathbb{R}$ where the integral is finite.

Notice that if $X, Y$ are independent then $M_{X+Y}(t) = M_X(t)M_Y(t)$. In particular, one can check that the moment generating functions behave consistently with the facts stated in Proposition 4.12 (i) and (ii).

**Example 6.9.** If $X$ is $\text{Bin}(n, p)$, then its moment generating function is $(1 + p(e^t - 1))^n$.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Moment generating function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal N(0,1)</td>
<td>$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$</td>
<td>$M(t) = e^{t^2/2}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$f(x) = \frac{1}{\mu} e^{-x/\mu}$</td>
<td>$M(t) = \frac{1}{1-e^{\mu t}}$</td>
</tr>
<tr>
<td>Uniform $U(-1,1)$</td>
<td>$f(x) = \frac{1}{2}$ for $-1 \leq x \leq 1$</td>
<td>$M(t) = \frac{1}{t} \sinh t$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$f(x) = 1/\Gamma(\alpha) x^{\alpha-1} e^{-x/\beta}$</td>
<td>$M(t) = (1 - \beta t)^{-\alpha}$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$</td>
<td>$M(t) = (1 - p + pe^t)^{n}$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$</td>
<td>$M(t) = e^{\lambda e^{-t}}$</td>
</tr>
<tr>
<td>Geometric</td>
<td>$\Pr(X = k) = p(1-p)^{k-1}$</td>
<td>$M(t) = e^{pe^t}$</td>
</tr>
</tbody>
</table>

**Theorem 6.28.** Suppose $M(s) < \infty$ on an open interval around $s = 0$. Then $M(s)$ has Taylor expansion $\sum_{k=0}^{\infty} \frac{s^k}{k!} E(X^k)$ with positive radius of convergence. In particular, $\mathbb{E}[X] = M'(0)$ and $E(X^2) = M''(0)$.

**Proof.** Suppose $M(s) < \infty$ for $|s| < \delta$.

We first note that $e^{sx} \leq e^{sx} e^{-sx}$. Indeed, if $sx > 0$ then $e^{sx} = e^{sx} e^{-sx}$. Similarly, if $sx < 0$ then $e^{sx} = e^{-sx} e^{sx}$.

Since

$$\frac{|sx|^k}{k!} \leq \sum_{k=0}^{\infty} \frac{|sx|^k}{k!} = e^{|sx|} \leq e^{sx} e^{-sx}$$

this shows that $E|X|^k < \infty$ for all $k > 0$. Furthermore, for $|s| < \delta$, random variables $S_n = \sum_{k=0}^{n} \frac{s^k}{k!} X^k$ are dominated by the integrable function $e^{sx} e^{-sx}$. Indeed, $|S_n| \leq \sum_{k=0}^{n} \frac{|s|^k}{k!} |X|^k \leq \sum_{k=0}^{\infty} \frac{|s|^k}{k!} |X|^k = e^{|s|X} \leq e^{sx} e^{-sx}$.

By dominated convergence, we can pass to the limit as $n \to \infty$ under the expectation sign. This gives $M(s) = \mathbb{E} e^{sx} = \lim_{n \to \infty} E S_n$.
By linearity of expected value, we have
\[ \mathbb{E}(S_n) = \sum_{k=0}^{n} \frac{s^k}{k!} \mathbb{E}X^k = \mathbb{E} \left( \sum_{k=0}^{n} \frac{s^k}{k!} X^k \right) \]

So the limit as \( n \to \infty \) is \( \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbb{E}(X^k) / k! \) and we get \( M(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbb{E}(X^k) / k! \), which is the Taylor series with radius of convergence \( R \geq \delta \). \( \square \)

**Corollary 6.29.** If \( M(s) < \infty \) for \( |s| < \delta \) then \( \mathbb{E}[e^{zX}] \) exists for \( z \in \mathbb{C} \) with \( -\delta < \Re z < \delta \). In particular, the characteristic function is \( \mathbb{E}\exp(itX) = M(it) \).

### Required Exercises

**Exercise 6.1.** Suppose that \( p, q, r \geq 1 \) are such that \( \frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1 \) and \( X, Y, Z \) are random variables such that \( \mathbb{E}|X|^p < \infty \), \( \mathbb{E}|Y|^q < \infty \), \( \mathbb{E}|Z|^r < \infty \). Use Hölder’s inequality to show that

\[ |\mathbb{E}(XYZ)| \leq \|X\|_p \|Y\|_q \|Z\|_r \]

**Solution:** Here are the steps: Let \( \frac{1}{s} = \frac{1}{q} + \frac{1}{r} = 1 - \frac{1}{p} \). Since \( p \geq 1 \), we see that \( s \geq 1 \). We will also need \( q' = q/s \geq 1 \), as \( r \geq 0 \) and \( r' = r/s \geq 1 \) as \( q \geq 0 \). We note that \( 1/q' + 1/r' = 1 \).

Apply Hölder inequality with \( p, s \) to get

\[ |\mathbb{E}(XYZ)| \leq \mathbb{E}(|X| \cdot |YZ|) \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}(|X|^s)^{1/s}) \]

Now \( \mathbb{E}(|X|^s) = \mathbb{E}(|X|^s|Y|^s) \) is a product, so we apply the Hölder inequality again, with \( q' = q/s \) and \( r' = r/s \). We get

\[ \mathbb{E}(|X|^s|Y|^s) \leq \mathbb{E}(|X|^{sq'})^{1/q'} \mathbb{E}(|Y|^{sr'})^{1/r'} = (\mathbb{E}|X|^q)^{s/q} (\mathbb{E}(|Y|^r)^{s/r} \]

**Exercise 6.2.** Suppose \( X_1, X_2, \ldots \) are independent square-integrable random variables with \( E(X_j) = 0 \) and \( E(X_j^2) = 1 \). Let \( S_n = X_1 + \cdots + X_n \). Prove that \( \frac{1}{n} E(|S_n|) \to 0 \). Hint: square-integrable random variables have second moments!

**Exercise 6.3.** Use Markov’s inequality to deduce that if \( \mathbb{E}[\exp X] < \infty \) then there exists \( C > 0 \) such that \( P(X > t) \leq Ce^{-t} \) for all \( t > 0 \). (This is another version of Exercise 5.17.)

**Exercise 6.4** (metric for convergence in probability). Define \( d(X, Y) = E \frac{|X-Y|}{1+|X-Y|} \). Prove that \( X_n \overset{p}{\to} X \) iff \( d(X_n, X) \to 0 \). (Note that this requires two proofs!)
Exercise 6.5. Suppose 
\[ P(X_n = n^3) = \frac{1}{n^2} \quad P(X_n = 0) = 1 - \frac{1}{n^2} \]
Show that \( X_n \to 0 \) with probability one but \( \lim_{n \to \infty} E(X_n) = \infty \).

Exercise 6.6. Suppose \( X_n \) are square-integrable random variables with uniformly bounded \( 1 + \delta \) moments \( E(|X_n|^{1+\delta}) \leq M \) for some \( \delta > 0 \). Prove that \( \{X_n\} \) is uniformly integrable, see Definition 6.3.

Exercise 6.7. Let \( X \geq 0 \) with \( \mathbb{E}X^2 < \infty \). Use Cauchy-Schwarz inequality to prove \( \mathbb{P}(Y > 0) \geq (\mathbb{E}(Y))^2/\mathbb{E}(Y^2) \).

Exercise 6.8. Prove that \( X \) is integrable if and only if for every \( \varepsilon > 0 \) there exists \( K \) such that \( \int_{|X|>K} |X(\omega)|dP < \varepsilon \). (Thus converse of Corollary 6.12 holds, too.)

Exercise 6.9. Use tail integration formula to show that if \( P(X > t) \leq Ce^{-t(1+\delta)} \) for some \( \delta > 0 \) then \( \mathbb{E}[\exp(X)] < \infty \).

Exercise 6.10. Suppose \( X,Y \) are independent with moment generating functions \( M_X(s) \) and \( M_Y(s) \). Show that the moment generating function of \( X + Y \) is \( M_X(s)M_Y(s) \). Hint: Proposition 6.27.

Exercise 6.11. Compute \( M(s) \) for binomial, Poisson, normal distributions, see Table 1 on page 80. Use the formula from Exercise 6.10 to verify that moment generating functions of sums of independent random variables are consistent with the facts listed in Proposition 4.12.

Exercise 6.12. Use Table 1 on page 80 and the formulas from Theorem 6.28 to verify formulas for the mean and variance given in the examples.

Exercise 6.13. Use Holder’s inequality to verify that \( \ln M(s) \) is a convex function.

Solution (H): We need to show that for \( 0 < \theta < 1 \) we have \( M(\theta s_1 + (1-\theta)s_2) \leq (M(\theta s_1))^\theta (M(s_2))^{1-\theta} \). Apply Holder’s inequality with \( 1/p = \theta, 1/q = (1-\theta) \) to the product \( \mathbb{E}(\exp(\theta s_1 X) \exp((1-\theta)s_2 Y)) \).

Exercise 6.14. Use Holder inequality to verify that the logarithm of the moment generating function is convex: \( \log M(\theta x + (1-\theta)y) \leq \theta \log M(x) + (1-\theta) \log M(y) \). Hint: it may be easier to first use Cauchy-Schwartz to verify this for \( \theta = 1/2 \).
Chapter 7

Product measure and Fubini’s theorem

This is based on [Billingsley, Section 18].

1. Product spaces

Suppose \((\Omega_1, \mathcal{F}_1)\) and \((\Omega_2, \mathcal{F}_2)\) are two probability spaces. In a product space \(\Omega = \Omega_1 \times \Omega_2\), a measurable rectangle is \(A_1 \times A_2\) where \(A_1 \in \mathcal{F}_1\) and \(A_2 \in \mathcal{F}_2\). We define \(\mathcal{F}_1 \otimes \mathcal{F}_2\) as the \(\sigma\)-field generated by measurable rectangles.

Example 7.1. \(B(\mathbb{R}) \otimes B(\mathbb{R}) = B(\mathbb{R}^2)\)

Theorem 7.1. (i) If \(E \in \mathcal{F}_1 \otimes \mathcal{F}_2\) then for each \(\omega_1 \in \Omega_1\) the set
\[
E_2(\omega_1) = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in E\}
\]
is in \(\mathcal{F}_2\).

Similarly, the other section: \(E_1(\omega_2) \in \mathcal{F}_1\).

(ii) if \(f : \Omega \to \mathbb{R}\) is measurable, then for each \(\omega_1 \in \Omega_1\) the function \(\omega_2 \mapsto f(\omega_1, \omega_2)\) is measurable.

Proof. Fix \(\omega_1 \in \Omega_1\). Consider the mapping \(T : \Omega_2 \to \Omega\) given by \(T(\omega_2) = (\omega_1, \omega_2)\). If \(E = A_1 \times A_2\) is a measurable rectangle, then \(T^{-1}(E)\) is either \(\emptyset\) or \(A_2\). So \(T\) is measurable, and \(T^{-1}(E) \in \mathcal{F}_2\) for all \(E \in \mathcal{F}_1 \otimes \mathcal{F}_2\).

Now \(g(\omega_2) = f \circ T(\omega_2)\) is a measurable mapping as a composition of measurable functions. \(\square\)

2. Product measure

Suppose \(P_1, P_2\) are probability measures on \(\mathcal{F}_1\) and \(\mathcal{F}_2\) respectively.

By Theorem 7.1, the function \(\omega_1 \mapsto P_2(E_2(\omega_1))\) is well defined.

The collection \(\mathcal{L}\) of subsets \(E\) of \(\Omega_1 \times \Omega_2\) for which this function is \(\mathcal{F}_1\)-measurable is a \(\lambda\)-system. The collection of measurable rectangles \(E = A_1 \times A_2\) is a \(\pi\)-system, and for such \(E\) the function is
\[ P_2(E_2(\omega_1)) = I_{A_1}(\omega_1)P_2(A_2), \] so it is measurable. Therefore, \( \omega_1 \mapsto P_2(E_2(\omega_1)) \) is measurable for all \( E \in \mathcal{F}_1 \otimes \mathcal{F}_2 \).

Since this is non-negative and bounded function, we can define

\[
P'(E) := \int_{\Omega_1} P_2(E_2(\omega_1))P_1(d\omega_1)
\]

and

\[
P''(E) := \int_{\Omega_2} P_1(E_1(\omega_2))P_2(d\omega_2)
\]

It is clear that both \( P' \) and \( P'' \) are probability measures. (Continuity follows from monotone convergence theorem for the integrals.)

We note that for measurable rectangles \( P'(A_1 \times A_2) = P''(A_1 \times A_2) = P_1(A_1)P_2(A_2) \).

Since the class of sets \( E \) for which \( P'' = P'' \) is a \( \lambda \)-system, this means that \( P' = P'' \). The common value is the product measure \( P = P_1 \otimes P_2 \).

Note that the above construction gives inductively a product measure on \( \Omega_1 \times \cdots \times \Omega_n \). In particular, it implies a “finite” version of Theorem 4.10.

**Theorem 7.2.** If \( F_1, F_2, \ldots, F_n \) are cumulative distribution functions then there exists a probability space \((\Omega, \mathcal{F}, P)\) and a sequence \( X_1, X_2, \ldots, X_n \) of independent random variables such that \( X_k \) has cumulative distribution function \( F_k \).

Somewhat more generally, if \( F_1 \) is a CDF on \( \mathbb{R}^m \), \( F_2 \) is a CDF on \( \mathbb{R}^n \), then there exists a CDF \( F \) on \( \mathbb{R}^{m+n} \) such that for \( x \in \mathbb{R}^m \), \( y \in \mathbb{R}^n \) we have \( F(x, y) = F_1(x)F_2(y) \).

### 3. Fubini’s Theorem

**Theorem 7.3.** If \( f : \Omega = \Omega_1 \times \Omega_2 \to \mathbb{R} \) is non-negative then the functions \( \omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)P_1(d\omega_1) \) and \( \omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)P_2(d\omega_2) \) are measurable and

\[
\int_{\Omega} f(\omega_1, \omega_2)d(P_1 \otimes P_2) = \int_{\Omega_2} \left( \int_{\Omega_1} f(\omega_1, \omega_2)P_1(d\omega_1) \right) P_2(d\omega_2) = \int_{\Omega_1} \left( \int_{\Omega_2} f(\omega_1, \omega_2)P_2(d\omega_2) \right) P_1(d\omega_1)
\]

If \( f \) is \( P_1 \otimes P_2 \)-integrable, then (7.3) holds.

**Sketch of the proof for** \( f \geq 0 \). For \( f = I_E \), formula (7.3) is just the definition of the product measure. By linearity, the same formula holds for simple \( f \). Now if \( f_n \uparrow f \) are simple and non-negative, then functions \( g_n(\omega_2) = \int_{\Omega_1} f_n(\omega_1, \omega_2)P_1(d\omega_1) \) are non-decreasing, non-negative, and converge to \( g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2)P_1(d\omega_1) \), so monotone convergence theorem (Theorem 6.4) gives (7.3). This ends the first part of the proof (the Tonelli’s theorem).

We omit the proof of the second part.\(^1\)

Fubini’s theorem is often a powerful computational tool. The following example illustrates its power and the value of the limit will be needed later on.

\(^1\)After the standard decomposition \( f = f^+ - f^- \) one needs to integrate over the set \( \Omega'_2 = \{ \omega_2 : \int_{\Omega_1} |f(\omega_1, \omega_2)|P_1(d\omega_1) < \infty \} \) which may be smaller than \( \Omega_2 \) but has \( P_2 \)-measure 1.
Example 7.2. The function $f(x, u) = \sin xu e^{-ux}$ is integrable on $(0, T) \times (0, \infty)$ as (by Tonelli’s theorem)

$$
\int_0^T \left( \int_0^\infty |e^{-ux} \sin x| du \right) dx = \int_0^T |\sin x| dx \leq T
$$

By Fubini’s theorem,

$$
\int_0^T \frac{\sin x}{x} dx = \int_0^T \left( \int_0^\infty e^{-ux} \sin x du \right) dx
\quad = \int_0^\infty \left( \int_0^T e^{-ux} \sin x dx \right) du \quad \text{(integrate by parts, twice)}
\quad = \int_0^\infty \frac{1}{1 + u^2} (1 - e^{-uT} (u \sin T + \cos T)) du \quad \text{(substitute } s = uT \text{ in the second integral)}
\quad = \pi/2 - \int_0^\infty e^{-s} \frac{s \sin T + T \cos T}{T^2 + s^2} ds
$$

Thus by the Lebesgue dominated convergence theorem

$$
\lim_{T \to \infty} \int_0^T \frac{\sin x}{x} dx = \frac{\pi}{2}
$$

(We will use this limit in the proof on page 116.)

Remark 7.4. Integrability with respect to the product measure is an important assumption that should always be checked! Durrett [Durrett] gives the following example: $f(x, y) = e^{-xy} - 2e^{-2xy}$. Then

$$
\int_0^1 \int_1^\infty f(x, y) dy dx = \int_0^1 \frac{e^{-x} - e^{-2x}}{x} dx > 0
$$

$$
\int_1^\infty \int_0^1 f(x, y) dx dy = \int_1^\infty \frac{e^{-2y} - e^{-y}}{y} dy < 0.
$$
7. Product measure and Fubini’s theorem

3.1. Integration by parts. If \( F, G \) have no common points of discontinuity in \((a, b)\) then

\[
\int_{(a, b)} G(x) dF(x) = F(b)G(b) - F(a)G(a) - \int_{(a, b)} F(x) G(dx)
\]

**Proof.** Write \((a, b) \times (a, b) = \Delta_+ \cup \Delta_-\) where \(\Delta_- = \{(x, y) : a < y \leq x \leq b\}\) and \(\Delta_+ = \{(x, y) : a < x \leq y \leq b\}\). The product measure is

\[
(F(b) - F(a))(G(b) - G(a)) = P((a, b) \times (a, b)) = P(\Delta_-) + P(\Delta_+) - P(\Delta_+ \cap \Delta_-)
\]

We note that \(\Delta_- \cap \Delta_+ = \{(x, x) : a < x \leq b\}\). By Fubini’s theorem

\[
P(\Delta_+ \cap \Delta_-) = \int_{(a, b)} (F[x]G(dx) = 0
\]

as \( F, G \) have no common point-mass atoms.

The formula is now a calculation, using

\[
P(\Delta_) = \int_{(a, b)} (G(x) - G(a)) F(dx)
\]

\[
P(\Delta_+) = \int_{(a, b)} (F(y) - F(a)) G(dy)
\]

\(\square\)

3.2. Tail integration formula. Formula (5.9) and its various generalizations are easy to derive from Fubini’s theorem: if \( X \geq 0 \) then \( X^p = \int_0^\infty pt^{p-1} I_{t<X} dt \) so

\[
E(X^p) = p \int_0^\infty t^{p-1} P(X > t) dt
\]

This formula holds true also in the non-integrable case - both sides are then \(\infty\).

3.3. Convolutions. Suppose \( X_1, X_2 \) are independent random variables with cumulative distribution functions \( F_1, F_2 \). Then the cumulative distribution function \( F_{X_1+X_2}(x) = P(X_1 + X_2 \leq x) \) is given by

\[
F_{X_1+X_2}(x) = \int_R F_1(x-u) F_2(du)
\]

**Proof.**

\[
P(X_1+X_2 \leq x) = \int_{\mathbb{R}^2} I_{u+v \leq x} P_1 \otimes P_2 (du, dv) = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} I_{u+v \leq x} P_1(du) \right) P_2(dv) = \int_{\mathbb{R}} F_1(x-v) P_2(dv)
\]

\(\square\)

**Example 7.3.** Suppose \( X_1 \) is uniform and \( X_2 \) is geometric. Then for \( x > 1 \) we get

\[
F_{X_1+X_2}(x) = \sum_{k=1}^\infty pq^{k-1} F_1(x-k) = pq[x]-1 F_1(x-[x]) + \sum_{k=1}^{[x]-1} pq^{k-1}
\]

\[
= pq[x]-1(x-[x]) + 1 - q[x]-1.
\]

So the density of \( X_1 + X_2 \) is \( f(x) = \sum_{k=1}^\infty pq^{k-1} f_{[k,k+1]}(x) \)
Required Exercises

Exercise 7.1. Suppose \( \lim_{n \to \infty} n^2 P(|X| > n) < \infty \). Prove that \( E|X| < \infty \)

Exercise 7.2. Suppose \( P(|X| > n) \leq \frac{1}{2^n} \) for all \( n \). Prove that there exists \( \delta > 0 \) such that \( E e^{\delta|X|} < \infty \).

Solution: We want to prove that \( \int_0^\infty e^{\delta t} P(|X| > t) dt < \infty \). By integral test for convergence, this is equivalent with \( \sum_{n=1}^{\infty} e^{\delta n} P(|X| > n) < \infty \). From the information provided, this series converges for any \( \delta < \ln 2 \).

Exercise 7.3. Use (7.5) to compute the cumulative distribution function for the sum of two independent uniform \( U(0,1) \) random variables. (Then differentiate to compute the density.)

Solution: \( F_{X+Y}(x) = \int F_X(x-u)F_Y(du) = \int_0^1 F(x-u)du \). Since \( F(x) = x_{1_{0<x<1}+1_{x>1}} \), we get

\[
F_{X+Y}(x) = \begin{cases} 
\int_0^x (x-u)du + \int_0^1 1_{x-u>1}du = \int_{\min\{1,x\}}^{\max\{0,x-1\}} (x-u)du + \int_0^{\min\{1,x-1\}} 1du & \text{for } x < 0 \\
\int_0^x (x-u)du + \int_0^1 1du = x^2/2 & \text{for } 0 \leq x < 1 \\
\int_0^{x-1} (x-u)du + \int_0^1 1du = 2x - 1 - x^2/2 & \text{for } 1 \leq x < 2 \\
\int_0^1 (x-u)du + \int_0^1 1du = 1 & \text{for } x \geq 2 
\end{cases}
\]

So the density is

\[
f(x) = \begin{cases} 
x & 0 < x < 1 \\
2 - x & 1 \leq x < 2 \\
0 & \text{otherwise} \end{cases}
\]

Exercise 7.4. Use (7.5) to compute the cumulative distribution function for the sum of two independent exponential \( (\lambda = 1) \) random variables. (Then differentiate to compute the density.)

Solution: Suppose \( x > 0 \).

\[
F(x-u)F(du) = \int_0^x (1 - e^{-(x-u)})1_{x-u>0}e^{-u}du = \int_0^x (1 - e^{-(x-u)})e^{-u}du \\
= \int_0^x e^{-u}du - e^{-x} \int_0^x 1du = 1 - e^{-x} - xe^{-x}
\]

So the density is \( f(x) = \frac{d}{dx} (1 - e^{-x} - xe^{-x}) = e^{-x} - e^{-x} + xe^{-x} = xe^{-x} \) for \( x > 0 \). (Case \( x < 0 \) is easier, as \( F_{X+Y}(x) = 0 \).)
Exercise 7.5. Use Fubini’s theorem (not tail integration formula from Theorem 6.22) to show that if $X \geq 0$ then

$$E \frac{1}{1+X} = 1 - \int_0^\infty \frac{1}{(t+1)^2} P(X > t) dt$$

Then re-derive the same result from Theorem 6.22.

Exercise 7.6. Use Fubini’s theorem (not tail integration formula from Theorem 6.22) to show that if $X \geq 0$ then

$$E e^X = 1 + \int_0^\infty e^t P(X > t) dt$$

Then re-derive the same result from Theorem 6.22.

Solution: Let’s prove a more general fact: if $f(x) = f(0) + \int_0^x g(u) du$ and $g \geq 0$ then

$$\mathbb{E}(f(X)) = f(0) + \int_0^\infty g(t) P(X > t) dt$$

**Proof.** $f(X(\omega)) = f(0) + \int_0^\infty 1_{u<X(\omega)} g(u) du$ Since $g \geq 0$ we can apply Fubini’s theorem:

$$\mathbb{E}(f(X)) = \int_\Omega f(X(\omega)) dP(\omega) = \int_\Omega \left( f(0) + \int_0^\infty 1_{u<X(\omega)} g(u) du \right) dP(\omega)$$

$$= f(0) + \int_0^\infty \int_\Omega 1_{u<X(\omega)} dP(\omega) g(u) du$$

$$= f(0) + \int_0^\infty P(u < X) g(u) du$$

□

Exercise 7.7. Suppose that $X,Y \geq 0$ are possibly dependent random variables and $p,q > 0$. Prove that

$$E(X^pY^q) = pq \int_0^\infty \int_0^\infty t^{p-1} s^{q-1} P(X > t, Y > s) dt ds$$

Additional Exercises

Exercise 7.8. Use Fubini’s Theorem to show that $\int_{\mathbb{R}} (F(x+a) - F(x)) dx = a$
Solution: To apply Fubini’s theorem, we assume $a \geq 0$ so that integrand will be non-negative.

$$
\int_{\mathbb{R}} (F(x+a) - F(x)) dx = \int_{\mathbb{R}} \int_{\Omega} (1_{X<x+a} - 1_{X<x}) dP dx
$$

$$
= \int_{\Omega} \int_{\mathbb{R}} (1_{X<x+a} - 1_{X<x}) dx dP = \int_{\Omega} \int_{X-a}^{X} 1 dx dP = \int_{\Omega} adP
$$

as $1_{X<x+a} - 1_{X<x} = \begin{cases} 
1 - 1 = 0 & x > X \\
1 - 0 = 1 & X - a < x < X \\
0 - 0 = 0 & x < X - a 
\end{cases}$

Case $a < 0$ is left for the interested reader to consider.

**Exercise 7.9.** Prove that for non-negative $X$ and $s > 0$ we have $E(\exp(sX)) = 1 + s \int_{0}^{\infty} e^{st} P(X > t) dt$.

(Solution: See solution of Exercise 7.6.)

**Exercise 7.10.** If $F$ is continuous, prove that $\int F(x)F(dx) = 1/2$. 
Sums of independent random variables

1. The strong law of large numbers

Theorem 8.1 (Etemadi). If $X_1, X_2, \ldots$ are pairwise independent identically distributed integrable random variables with mean $m$ then $\frac{1}{n}S_n \to m$ with probability one.

For a completely elementary (but not simple) proof of this result is [Billingsley, Theorem 22.1]. Instead we will exhibit several proof techniques which can be applied to prove laws of large numbers under various sets of assumptions. The proof techniques rely on Markov inequality, Borel-Cantelli lemma, decomposition into positive/negative parts, and truncation.

The following is known as a weak law of large numbers, and the proof is an exercise.

Theorem 8.2. If $X_1, X_2, \ldots$ are pairwise independent with the same mean and uniformly bounded second moments, then $\frac{1}{n}S_n \overset{P}{\to} m$.

Proof. Compute $\text{Var} \frac{1}{n}S_n$. □

Theorem 8.3. If $X_1, X_2, \ldots$ are quadruple-independent with the same mean $m$ and with uniformly bounded fourth moments then $\frac{1}{n}S_n \to m$ with probability one.

Proof. Without loss of generality we can assume $m = 0$. (Replace $X_n$ by $X_n - m$.) We will use Borel-Cantelli lemma to verify that for every $\varepsilon > 0$, $P(\frac{1}{n}|S_n| \geq \varepsilon \ i.o.) = 0$. To do so, we use Markov’s

---

1This proof assumes that $X_1, X_2, \ldots$ have the same distribution. You should figure out what needs to be modified if the distributions are not the same!!!
inequality,

$$P \left( \frac{1}{n} |S_n| \geq \varepsilon \right) \leq \frac{\mathbb{E}[S_n^4]}{\varepsilon^4 n^4}$$

We note that by Laypunov (or Cauchy-Schwartz) inequality $E(X^2_j) \leq \sqrt{EX^4_j} < \infty$, so for identically distributed random variables

$$\mathbb{E}[S^4_n] = \sum_{j_1,j_2,j_3,j_4=1}^n \mathbb{E}[X_{j_1}X_{j_2}X_{j_3}X_{j_4}] = n\mathbb{E}(X^4_1) + 3n(n-1)(\mathbb{E}[X^2_1])^2 \leq Cn^2$$

Thus $\sum_n P(\frac{1}{n} |S_n| \geq \varepsilon) < \infty$. By Borel-Cantelli (Theorem 3.8) $P(\frac{1}{n} |S_n| > \varepsilon \text{ i.o.}) = 0$, ending the proof. (See discussion of convergence with probability one in the proof of Proposition 4.14.)

**Theorem 8.4.** If $X_1, X_2, \ldots$ are pairwise independent identically distributed square-integrable random variables with mean $m$ then $\frac{1}{n}S_n \to m$ with probability one.

**Proof.** The proof of Theorem 8.2 shows that $\frac{1}{m^2}S_{n^2} \to m$ with probability one. Writing $X_j = X^+_j - X^-_j$, when random variables have the same distribution, without loss of generality we may assume $X_j \geq 0$. Then for $n \in \mathbb{N}$ choose $k = k(n)$ such that $k^2 \leq n < (k+1)^2$. Notice that $n/k^2 \to 1$.

Since we know that $\frac{1}{k^2}S_{k^2} \to m$ with probability one, we have

$$\frac{S_{k^2}}{n} \leq \frac{S_n}{n} \leq \frac{S_{(k+1)^2}}{n}$$

And

$$\frac{S_{k^2}}{n} = \frac{S_{k^2}}{k^2} \frac{k^2}{n} \to m, \quad \frac{S_{(k+1)^2}}{n} = \frac{S_{(k+1)^2}}{(k+1)^2} \frac{(k+1)^2}{n} \to m$$

So by the squeeze theorem, $\frac{S_n}{n} \to m$.

One can reduce moment assumptions by a suitable use of truncation.

---

(Omitted in 2019) The following result assumes less than Theorem 8.3, but more than Theorem 8.1. It is harder to prove, but still easier to prove than Theorem 8.1.

**Theorem 8.5.** If $X_1, X_2, \ldots$ are pairwise independent identically distributed random variables with mean $m$, and $E(|X_1|^{3/2}) < \infty$, then $\frac{1}{n}S_n \to m$ with probability one.
1. The strong law of large numbers

**Proof.** (Omitted in 2019) The main steps in the proof are the same as in the proof of Theorem 8.1: are reduction to non-negative case, truncation, Borel-Cantelli lemma, and the use of subsequences. But the technicalities are somewhat simpler due to the stronger moment assumption. Since $m = m^+ - m^-$ and $X_j = X_j^+ - X_j^-$ without loss of generality we can assume that $X_j \geq 0$.

Next, we introduce truncation $X_k = X_k' + X_k''$ where $X_k' = X_k I_{X_k \leq k}$ and denote $S_n' = \sum_{k=1}^n X_k'$.

Since $X_k$ are identically distributed and integrable,

\[
\sum_{k=1}^\infty P(X_k \neq X_k') = \sum_{k} P(|X_k| > k) = \sum_{k} P(|X_1| > k) < \int_0^\infty P(|X_1| > t) dt = E(|X_1|) < \infty
\]

Therefore by the Borel-Cantelli Lemma $P(X_k \neq X_k' \ i.o.) = 0$. This implies that $\frac{1}{n}S_n - \frac{1}{n}S_n' \to 0$ with probability one.

(Omitted in 2019) Next we compute

\[
\text{Var}(S_n') = \sum_{k=1}^n \text{Var}(X_k') \leq \sum_{k=1}^n E(X_k^2 I_{|X_k| \leq k}) = \sum_{k=1}^n E(X_k^2 I_{|X_k| \leq k}) \leq nE(X_1^2 I_{|X_1| \leq n}).
\]

We apply inequality (8.1) to subsequence $\frac{1}{n^2}S_n'$. We get

\[
\sum_{n=1}^\infty \text{Var}(S_{n^2}/n^2) \leq \sum_{n=1}^\infty \frac{n^2}{n^2} E(X_1^2 I_{|X_1| \leq n^2}) = E \left( \frac{|X_1|^2}{n^2} \sum_{n^2 \geq |X_1|} \frac{1}{n^2} \right) \\
\leq C + \int_{|X_1| \geq 4} \left( \frac{|X_1|^2}{n^2} \sum_{n^2 \geq |X_1|} \frac{1}{n(n-1)} \right) \frac{1}{|X_1|-1} dP \\
= C + \int_{|X_1| \geq 4} \left( \frac{|X_1|^2}{\sqrt{|X_1|-1}} \right) dP \leq C + 2E(|X_1|^{3/2}).
\]

Here we use the bound $\frac{\sqrt{2\pi}}{\sqrt{2\pi-1}} \leq \frac{\sqrt{x}}{\sqrt{x-1}} = 2$ for $x \geq 4$. Therefore, Chebyshev’s inequality implies that

\[
\sum_{n} P\left( \frac{1}{n^2} |S_{n^2}' - E(S_{n^2}')| > \varepsilon \right) < \infty
\]

By another application of Borel-Cantelli, we see that $\frac{1}{n^2} \left( S_{n^2}' - E(S_{n^2}') \right) \to 0$ with probability one.

Now we note that $\frac{1}{n}E(S_n') = \frac{1}{n} \sum_{k=1}^n m_k$ where $m_k = E(X_1 I_{|X_1| \leq k}) \to m$ by Lebesgue’s dominated convergence theorem (Theorem 6.9). By Cesaro’s theorem the sequence $\frac{1}{n}E(S_n')$ also converges to $m$. 
(8.2) \[ \frac{1}{n}S_n - m = \frac{1}{n}(S_n - S_n') + \frac{1}{n}(S_n' - E(S_n')) + \left( \frac{1}{n} \sum_{k=1}^{n} m_k - m \right) \]

From (8.2) we therefore deduce that \( \frac{1}{n}S_n \to m \) with probability one.

To conclude the proof, we note that every number \( n \) lies between two perfect squares. That is, we introduce the sequence \( k_n = \lfloor \sqrt{n} \rfloor \) so that \( k_n^2 \leq n \leq (k_n + 1)^2 \).

Notice that \( k_n = k_{n+1} \) is possible, but \( k_n \to \infty \) eventually.

By the previous part of the proof, we know that \( \frac{1}{k_n^2}S_{k_n^2} \to m \) and \( \frac{1}{(k_{n+1})^2}S_{(k_{n+1})^2} \to m \) with probability one.

We now use \( X_j \geq 0 \) to infer that

\[ S_{k_n^2} \leq S_n \leq S_{(k_{n+1})^2} \]

so \( \frac{1}{k_n^2}S_n \) is between \( \frac{k_n^2}{(k_{n+1})^2} \frac{1}{k_n^2}S_{k_n^2} \) and \( \frac{(k_{n+1})^2}{k_n^2} \frac{1}{(k_{n+1})^2}S_{(k_{n+1})^2} \).

Since \( \frac{k_n}{k_{n+1}} \to 1 \) as \( k \to \infty \), the result follows.

\[ \square \]

2. Kolmogorov’s zero-one law

(Omitted in 2019)

**Theorem 8.6.** Suppose \( X_1, X_2, \ldots \) are independent and \( A \in \mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots) \). Then \( P(A) = 0 \) or \( P(A) = 1 \).

**Proof.** Consider \( \mathcal{F}_0 = \bigcup_{n=1}^{\infty} \sigma(X_1, \ldots, X_k) \). We first check that \( \mathcal{F}_0 \) is a field, and that it generates \( \sigma(X_1, X_2, \ldots) \).

Next suppose \( A \in \mathcal{T} \). Then \( A \in \sigma(X_{k+1}, \ldots) \) for every \( k \), so \( A \) is independent of \( B \in \sigma(X_1, \ldots, X_k) \). So \( A \) is independent of \( \mathcal{F}_0 \). Since \( \mathcal{F}_0 \) is a \( \pi \)-system, hence \( A \) is independent of \( \sigma(\mathcal{F}_0) \). But that means that \( P(A)P(A) = P(A) \). \( \square \)

3. Kolmogorov’s Maximal inequality and its applications

**Theorem 8.7** (Kolmogorov’s maximal inequality). Suppose \( X_1, X_2, \ldots, X_n \) are independent with mean 0 and finite variance. For \( t > 0 \),

\[
(8.3) \quad P(\max_{1 \leq k \leq n} |S_k| \geq t) \leq \frac{\text{Var}(S_n)}{t^2}
\]

**Proof.** We have a sequence of random variables \( |S_1|, |S_2|, \ldots, |S_n| \) and we want to estimate the probability that at least one of them exceeds level \( t \). The trivial estimate \( P(\max_{1 \leq k \leq n} |S_k| \geq t) \leq \sum_{k=1}^{n} P(|S_k| > t) \) is not accurate enough due to multiple overlaps, so we must decompose the event \( \{\max_{1 \leq k \leq n} |S_k| > t\} \) more carefully into disjoint sets. The trick is to look at where the first crossing of level \( t \) occurs. Consider the (disjoint) events

\[ A_k = \{|S_1| < t, \ldots, |S_{k-1}| < t, |S_k| \geq t\}. \]

Clearly, \( P(\max_{1 \leq k \leq n} |S_k| \geq t) = \sum_{k=1}^{n} P(A_k) \).
Since the events are disjoint, we have

\[
E(S_n^2) \geq \int_{\max_{1 \leq k \leq n} |S_k| \geq t} S_n^2 dP = \sum_{k=1}^{n} \int_{A_k} S_n^2 dP = \sum_{k=1}^{n} \int_{A_k} (S_n - S_k + S_k)^2 dP = \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)^2 + S_k^2 dP + 2 \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)S_k dP
\]

\[
\geq \sum_{k=1}^{n} \int_{A_k} S_k^2 dP + 2 \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)S_k dP
\]

\[
\geq t^2 \sum_{k=1}^{n} P(A_k) + 2 \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)S_k dP = t^2 P(\max_{1 \leq k \leq n} |S_k| \geq t) + 2 \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)S_k dP
\]

This is what we want, if we can show that the last sum vanishes. And this is indeed the case:

\[
\int_{A_k} (S_n - S_k)S_k dP = \int_{\Omega} (S_n - S_k)S_k I_{A_k} dP = E(S_n - S_k)E(S_k I_{A_k}) = 0
\]

as the event \(A_k\) is \(\sigma(X_1, \ldots, X_k)\)-measurable, \(S_k\) is also \(\sigma(X_1, \ldots, X_k)\)-measurable, but \((S_n - S_k) = X_{k+1} + \cdots + X_n\) is \(\sigma(X_{k+1}, X_{k+2}, \ldots, X_n)\)-measurable. This shows that random variables \(U = (S_n - S_k)\) and \(V = S_k I_{A_k}\) are independent, and we can apply the formula \(E(UV) = E(U)E(V)\).

**Theorem 8.8** (Kolmogorov’s one-series theorem). Suppose that \(\{X_n\}\) is an independent sequence, \(E(X_n) = 0\) and \(\sum_n \text{Var}(X_n) < \infty\). Then the series \(\sum_{n=1}^{\infty} X_n\) converges with probability one.

**Proof.** Denote \(S_n = \sum_{k=1}^{n} X_k\). From (8.3) we have

\[
P(\max_{1 \leq k \leq r} |S_{n+k} - S_n| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sum_{k=1}^{r} \text{Var}(X_{n+k})
\]

Taking \(r \to \infty\), (these are increasing events)

\[
P(\sup_k |S_{n+k} - S_n| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sum_{k=1}^{\infty} \text{Var}(X_{n+k})
\]

Since the variances converge,

\[
(8.4) \quad \lim_{n \to \infty} P(\sup_{k \geq 1} |S_{n+k} - S_n| > \varepsilon) = 0
\]

Now look at \(A_{n,\varepsilon} = \{\sup_{j, k \geq n} |S_j - S_k| > 2\varepsilon\}\). Notice that these are decreasing events, \(A_{n,\varepsilon} \supset A_{n+1,\varepsilon}\). Indeed,

\[
\sup_{j, k \geq n} |S_j - S_k| \geq \sup_{j, k \geq n+1} |S_j - S_k|
\]

We have

\[
A_{n,\varepsilon} = \{\sup_{j, k \geq n} |S_j - S_n + S_n - S_k| > 2\varepsilon\} \subset \sup_{j, k \geq n} |S_j - S_n| + |S_n - S_k| > 2\varepsilon
\]

\[
= \sup_{j \geq n} |S_j - S_n| + \sup_{k \geq n} |S_n - S_k| > 2\varepsilon \subset \{\sup_{j \geq n} |S_j - S_n| > \varepsilon\} \cup \{\sup_{k \geq n} |S_k - S_n| > \varepsilon\}
\]

\[
= \{\sup_{j \geq 1} |S_{n+j} - S_n| > \varepsilon\} \cup \{\sup_{k \geq 1} |S_{n+k} - S_n| > \varepsilon\}
\]
Since \( P(A \cup B) \leq P(A) + P(B) \), we see that (8.4) implies that \( P(A_{n, \varepsilon}) \to 0 \). This shows that \( P(\cap_n A_{n, \varepsilon}) = 0 \). So (by taking union over all rational \( \varepsilon > 0 \)) we see that

\[
P(\exists \varepsilon > 0 \forall n \sup_{j, k \geq n} |S_j - S_k| > 2\varepsilon) = 0
\]

That is,

\[
P(\forall \varepsilon > 0 \exists n \exists j, k \geq n |S_j - S_k| < 2\varepsilon) = 1
\]

This means that there is \( \Omega_0 \subseteq \Omega \) of probability one such that for all \( \omega \in \Omega_0 \) the sequence of numbers \( \{S_k(\omega)\}_{k \in \mathbb{N}} \) is a Cauchy sequence, i.e., \( \lim_{n \to \infty} S_n(\omega) \) exists.

**Corollary 8.9** (Kolmogorov’s two series theorem). Suppose that \( \{X_k\} \) is independent and that the following two series converge:

\[
\sum_n E(X_n) \text{ converges}, \quad \sum_n \text{Var}(X_n) < \infty
\]

Then \( \sum_n X_n \) converges.

**Proof.** \( \sum_{k=1}^n X_k = \sum_{k=1}^n (X_k - E(X_k)) + \sum_{k=1}^n E(X_k) \), so we get the sum of two convergent series. \( \square \)

The following corollary is useful when dealing with series of random variables without variances.

**Theorem 8.10** (Kolmogorov’s three series theorem). Suppose that \( \{X_k\} \) is independent and that for some positive \( c > 0 \) the following three series converge:

\[
\sum_n P(|X_n| > c) < \infty, \quad \sum_n E(X_n I_{|X_n| \leq c}) \text{ converges}, \quad \sum_n \text{Var}(X_n I_{|X_n| \leq c}) < \infty
\]

Then \( \sum_n X_n \) converges.

**Proof.** Let \( X'_n = X_n I_{|X_n| \leq c} \) denote the truncated random variables. Define \( m_n = E(X'_n) \). By Theorem 8.8, \( \sum_n (X'_n - m_n) \) converges with probability one. Since \( \sum_n m_n \) converges, therefore \( \sum_n X'_n \) converges. Now we note that

\[
\sum_n P(X_n \neq X'_n) = \sum_n P(|X_n| > c) < \infty
\]

so by Borel-Cantelli’s Lemma, \( P(X_n \neq X'_n \text{ i.o.}) = 0 \). Thus the series \( \sum_n X_n \) converges. \( \square \)

**Remark 8.11.** If \( \sum_n X_n \) converges with probability one, then \( X_n \to 0 \) with probability one. So Borel Cantelli Lemma implies that \( \sum_n P(|X_n| > c) < \infty \) for any \( c > 0 \). One can show that condition (8.6) then holds, so Theorem 8.10 is in fact an equivalence.

**Example 8.1.** It is well known that the harmonic series \( \sum_n \frac{1}{n} \) diverges while the alternating series \( \sum_n (-1)^n n \) converges. It is somewhat comforting to know that the latter is more typical: Suppose \( \varepsilon_k = \pm 1 \) is an infinite sequence of signs such that every \( n \)-tuple of \( 2^n \) possible signs \((\pm 1, \pm 1, \ldots, \pm 1)\) is equally likely. Then the series \( \sum_n \varepsilon_n \) converges with probability one.

**Proof.** It is easy to see that \( \varepsilon_1, \varepsilon_2, \ldots \) are independent random variables with mean 0 and variance 1, compare Example 4.5. So with \( X_n = \varepsilon_n / n \) the result follows from Theorem 8.8. \( \square \)

**Example 8.2.** Suppose \( \{X_k\} \) is a sequence of independent uniform random variables \( U(-a_k, a_k) \).

(i) If \( \sum a_k < \infty \) then \( \sum |X_k| \) converges with probability one. (This is trivial!!)
(ii) if \( \sum a_k^2 < \infty \) then \( \sum X_k \) converges with probability one.

**Example 8.3.** Suppose \( \{X_k\} \) is a sequence of independent standard normal random variables. If \( \sum a_k^2 < \infty \) then \( \sum X_k \) converges with probability one. In particular, \( W_t = \sum_{n=1}^{\infty} \frac{e^{itn}}{n} X_n \) converges.

**Example 8.4.** Suppose \( X_n \) are i.i.d. Cauchy with density \( \frac{1}{\pi (1+x^2)} \). Then \( \sum a_n X_n \) converges if and only if \( \sum |a_n| < \infty \). Indeed, since \( X_n \) and \( -X - n \) have the same law, and we can skip the terms where \( a_n = 0 \), so without loss of generality we may assume \( a_n > 0 \). Then

\[
P(|a_n X_n| > 1) = \frac{1}{\pi} \int_{1/a_n}^{\infty} \frac{dx}{1+x^2} = \frac{1}{\pi} (\pi/2 - \arctan 1/a_n)
\]

Noting that

\[
\lim_{x \to \infty} \frac{\pi/2 - \arctan x}{1/x} = \lim_{x \to \infty} -\frac{1}{1+x^2} = 1
\]

we see that if \( a_n \to 0 \) then

\[
\lim_{n \to \infty} \frac{P(|a_n X_n| > 1)}{a_n} = 1
\]

This shows that the convergence if the series \( \sum_n P(|a_n X_n| > 1) \) is equivalent to \( \sum_n |a_n| < \infty \).

Now if \( \sum |a_n| < \infty \) then (8.6) holds: \( E(a_n X_n I_{|a_n X_n| \leq 1}) = 0 \) because the law of \( X_n \) is symmetric, and

\[
\sum_n E(a_n^2 X_n^2 I_{|a_n X_n| \leq 1}) = \frac{1}{\pi} \sum_n a_n^2 \int_{1/a_n}^{\infty} \frac{x^2}{1+x^2} dx \leq \frac{1}{\pi} \sum_n a_n^2 \int_{-1/a_n}^{1/a_n} dx = \frac{2}{\pi} \sum_n a_n < \infty
\]

**Remark 8.12** (Levy’s theorem). It is known that for independent random variables \( \sum_n X_n \) converges in distribution iff it converges in probability iff it converges with probability one. See [Varadhan, Theorem 3.9] or (for one implication) [Billingsley, Theorem 22.7]. This result holds true also in infinite dimensional setting (Ito-Nisio)

3.0.1. **Kolmogorov’s Strong Law of Large Numbers.** Here is another proof of the strong law of large numbers - this proof uses joint independence, and second moments.

**Corollary 8.13.** Suppose \( \{X_k\} \) is independent with the same mean \( m \) and uniformly bounded (finite) variances. Then \( \frac{1}{n} S_n \to m \) with probability one.

**Proof.** Subtracting \( m \) if necessary, without loss of generality we assume that \( m = 0 \). Since \( Var(\frac{1}{n} X_n) = \sigma^2/n^2 \) and the series \( \sum_n 1/n^2 \) converges, by Theorem 8.8, \( \sum_n \frac{1}{n} X_n \) converges with probability one.

We now use a Lemma for sequences of numbers.

**Lemma 8.14** (Kronecker’s Lemma). If the series \( \sum_n x_n/n \) converges then \( \frac{1}{n} (x_1 + \cdots + x_n) \to 0 \).
8. Sums of independent random variables

(Omitted in 2019)

**Proof.** Write \( s_n = x_1 + \cdots + x_n \) and \( t_n = \sum_{k=1}^{n} x_k/k \).

We have \( x_k = k(t_k - t_{k-1}) \) so

\[
\frac{1}{n} s_n = \frac{1}{n} \sum_{k=1}^{n} k(t_k - t_{k-1}) = \frac{1}{n} \sum_{k=1}^{n} \sum_{j=1}^{k} (t_k - t_{k-1}) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=j}^{n} (t_k - t_{j-1}) = t_n - \frac{1}{n} \sum_{j=1}^{n} t_{j-1}
\]

Now \( t_n \to t_\infty \) by assumption and the average must\(^3\) have the same limit.

So \( \frac{1}{n} s_n \to t_\infty - t_\infty = 0. \)

\( \Box \)

(Omitted in 2019)

**Corollary 8.15** (Kolmogorov’s strong law of large numbers). Suppose \( \{X_n\} \) is independent identically distributed with mean \( m \). Then \( \frac{1}{n} S_n \to m \) with probability one.

**Proof.** As previously we consider \( S_n = \sum_{k=1}^{n} X_k \) where \( X_k = X_k I_{|X_k| \leq k} \). As previously, \( (S_n - S_n')/n \to 0 \), and \( E(S_n')/n \to m \), so we only need to show that \( (S_n' - E(S_n'))/n \to 0 \).

To prove that \( (S_n' - E(S_n'))/n \to 0 \) we use Kronecker’s Lemma 8.14. That is, we verify that \( \sum_{k} (X_k' - E(X_k'))/k \) converges with probability one. To apply Kolmogorov’s convergence criterion (Theorem 8.8) we only need to verify that \( \sum_{n=1}^{\infty} \text{Var}(X_k')/k^2 < \infty \).

The main estimate is similar to the one we already saw in the proof of Theorem 8.1.

\[
\sum_{k=1}^{\infty} \frac{\text{Var}(X_k')}{k^2} \leq \sum_{k=1}^{\infty} k \frac{1}{k^2} E(X_k^2 I_{|X_k| \leq k}) = E \left( \frac{X_1^2}{|X_1|} \sum_{k=1}^{\infty} \frac{1}{k^2} \right) = \int_{|X_1| \leq 1} \frac{X_1^2}{|X_1|} \sum_{k=1}^{\infty} \frac{1}{k^2} \leq \frac{\pi^2}{6} + \int_{|X_1| \geq 1} \left( \frac{|X_1|^2}{|X_1|^2 + \int_{|X_1|}^{\infty} \frac{1}{t^2} dt} \right) dP
\]

Here we used the fact that if \( n \geq 1 \) \( \leq |X_1| \leq n \) then \( 1/n^2 \leq 1/|X_1|^2 \), so

\[
\sum_{k \geq |X_1|} \frac{1}{k^2} \leq \frac{1}{n^2} + \sum_{k=n+1}^{\infty} \frac{1}{k^2} \leq \frac{1}{n^2} + \int_{n}^{\infty} \frac{1}{t^2} dt \leq \frac{1}{|X_1|^2} + \int_{|X_1|}^{\infty} \frac{1}{t^2} dt.
\]

\( \Box \)

4. Etemadi’s inequality and its application
Exercise 8.1. Suppose \( \{X_k\} \) are independent uniform \( U(0, 1) \) random variables. Prove that \( \max_{1 \leq k \leq n} X_k \to 1 \) with probability one.

Exercise 8.2. Suppose \( \{X_k\} \) are independent uniform \( U(0, 1) \) random variables. Prove that \( X_1 X_2 \ldots X_n \to 0 \) with probability one. Hint: There are many proofs, but perhaps the easiest uses Markov’s inequality. (This is Exercise 4.16 that was solved differently!)

\[ \text{Solution: In fact, the series } \sum_{n=1}^{\infty} X_1 X_2 \ldots X_n \text{ converges, as the terms are non-negative and the series of expected values converges.} \]

Exercise 8.3. Suppose \( \{X_k\} \) are independent uniform \( U(0, 1) \) random variables. Prove that the geometric means \( \{\sqrt[n]{X_1 X_2 \ldots X_n}\}_{n \in \mathbb{N}} \) converge with probability one. (And find the limit!)

Required Exercises
Exercise 8.4. Let $X_1, X_2, \ldots$ be identically distributed random variables with finite second moments. Show that $nP(|X_1| > \varepsilon \sqrt{n}) \to 0$ and $n^{-1/2} \max_{k \leq n} |X_k| \xrightarrow{P} 0$. Hint: Second moments imply that for every $\varepsilon > 0$ one can find $K$ such that $EX_1^2 I_{|X_1| > K} < \varepsilon$.

Exercise 8.5. Suppose $\{X_n\}$ is independent identically distributed and integrable. Prove that although we omitted the proof of the theorem you need here, so this solution is "less complete"!

Hint #1: Theorem 8.8 both series $\sum_{X \geq 0}$ and $\sum_{X > 0}$ are independent exponential random variables such that $\sum_{X \geq 0}$, $\sum_{X > 0}$ and $\sum_{X = 0}$ are constructed iteratively by the following procedure: $X_1$ is uniform $U(0, 1)$, and for $n \geq 1$, $X_{n+1}$ has uniform distribution on $(0, X_n)$. Show that $\frac{1}{n} \log X_n$ converges with probability one.

Exercise 8.6. Suppose $\{X_n\}$ is a sequence of independent identically distributed exponential random variables. Prove that although $P(\log X_1 \to 0) = 1$, we have $P(\log X_n \to 0) = 0$.

Exercise 8.7. Suppose $X_n$ are constructed iteratively by the following procedure: $X_1$ is uniform $U(0, 1)$, and for $n \geq 1$, $X_{n+1}$ has uniform distribution on $(0, X_n)$. Show that $\frac{1}{n} \log X_n$ converges with probability one and find the limit.

Exercise 8.8. Suppose $X_n$ are independent exponential with parameter $\lambda_n > 0$. For which $\lambda_n$ we have $X_n \to 0$ in mean? In mean square? In probability? With probability one?

Exercise 8.9. Suppose $X_n$ are independent random variables such that $X_n \to X$ with probability one. Show that $X$ cannot take two different values.

Exercise 8.10. Suppose $\{X_n\}$ are independent with mean 0 and finite variance. Let $S_n = \sum_{k=1}^{n} X_1X_2 \ldots X_k$. Adapt the proof of Kolmogorov’s maximal inequality to estimate $P(\max_{1 \leq k \leq n} |S_k| \geq t)$.

Exercise 8.11. Suppose $\{X_n\}$ are independent identically distributed with mean 0 and finite variance. Show that the series $\sum_{n=1}^{\infty} \frac{1}{n} X_n X_{n+1}$ converges almost surely.

Solution: Let $Y_n = \frac{1}{n} X_n X_{n+1}$. Then $E(Y_n) = 0$ and $E(Y_n^2) = \sigma^4/n^2$. Since $Y_1, Y_3, \ldots, Y_{2n+2}, \ldots$ are independent and $Y_2, Y_4, \ldots, Y_{2m+2}, \ldots$ are independent, by Theorem 8.8 both series $\sum_{k=1}^{n} Y_{2k+1}$ and $\sum_{k=1}^{n} Y_{2k}$ converge with probability one. If the first series converges on $\Omega_1$ and the second converges on $\Omega_2$, then the sum of both series converges on $\Omega_1 \cap \Omega_2$, and this is the set of probability one: $P(\Omega_1 \cap \Omega_2) = 1 - P(\Omega_1^c \cup \Omega_2^c) \geq 1 - P(\Omega_1^c) - P(\Omega_2^c) = 1$.

Let $S_n = \sum_{k=1}^{n} Y_k$. The above argument shows that partial sums $\sum_{k=0}^{n} (Y_{2k+1} + Y_{2k+2}) = S_{2n+2}$ converge to some (random) limit $S$. To prove convergence of the series, we need to show more: we need to show that all partial sums $S_n$ converge to the same limit. To complete the proof, we note that $S_n = S_{2m}$ or $S_n = S_{2m+1} = S_{2m} + Y_{2m+1}$. To complete the proof, we prove that $S_{2m+1} \to S$ by showing that $Y_{2m+1} \to 0$ with probability one. The latter follows from convergence of the series. Or from Borel-Cantelli: $P(|Y_n| > \varepsilon) \leq \frac{\sigma^4}{\varepsilon^2 n^2}$ so $\sum_{n} P(|Y_n| > \varepsilon) < \infty$. (Hence for any $\varepsilon > 0$ we have $P(|Y_n| > \varepsilon \text{ i.o.}) = 0$.)

\[4\text{That is, } P(X_n \leq x) = 1 - e^{-x} \text{ for } x > 0\]
Exercise 8.12. Suppose \( \{X_n\} \) is independent identically distributed square-integrable with mean \( E(X_k) = 0 \). Let \( \{c_n\} \) be a bounded sequence of numbers. Modify the proof of the law of large numbers (which one?) to show that \( \frac{1}{n} \sum_{k=1}^{n} c_k X_k \to 0 \) with probability one.

Exercise 8.13. Suppose \( X_n \) are independent identically distributed integrable with symmetric distribution: \( X_1 \) has the same law as \( -X_1 \). Prove that the series \( \sum \frac{1}{n} X_n \) converges with probability one. Hint: Theorem 8.10 is in fact an equivalence.

Solution: We verify condition (8.6) with \( c = 1 \).

\[
\sum_{n} P \left( \frac{|X_n|}{n} > 1 \right) = \sum_{n} P(|X_1| > n) \leq \int_{0}^{\infty} P(|X| > t)dt = E|X| < \infty
\]

By symmetry, \( E(X_n I_{|X_n|\leq n}) = 0 \), so the second series converges. The verification of the last condition is based on Fubini’s theorem as in the proof of Kolmogorov’s SLLN. This proof uses the fact that \( \sum_{n \geq x} \frac{1}{n^2} \sim C/x \) for \( x > 1 \), which is re-proved within the argument.

\[
\sum_{n} \frac{1}{n^2} E(X_n^2 I_{|X_n|\leq n}) = E(\sum_{n} \frac{1}{n^2} X_n^2 I_{|X_n|\leq n})
= E \left( X^2 \sum_{n \geq |X|} \frac{1}{n^2} \right) \leq E(X^2 I_{|X|\leq 1}) + E \left( X^2 I_{|X|>1} \int_{1}^{|X|} \frac{du}{u^2} \right)
= E(X^2 I_{|X|\leq 1}) + E \left( X^2 I_{|X|>1} \left( \frac{1}{|X|} - 1 \right) \right)
\leq E(X^2 I_{|X|\leq 1}) + E \left( X^2 \frac{1}{|X|} \right) \leq 1 + E|X| < \infty
\]

Exercise 8.14. Modify our proof of Theorem 8.1 under extra moments to show that \( \frac{1}{n} S_n \to m \) with probability one under additional moment assumption that \( E(|X_1|^{1+\delta}) < \infty \) for some \( \delta > 0 \). Hint: Consider the subsequence \(^5 \frac{1}{n^\theta} S_{n^\theta} \).

Exercise 8.15. Modify our proof of Theorem 8.1 under extra moments to show that under the assumption \( E(|X_1|) < \infty \) we have \( \frac{1}{n^\theta} S_{2^n} \to m \) with probability one. Hint: An important step is to use (8.1) to show that \( \sum \text{Var}(S_{2^n})/4^n < \infty \), see [Billingsley, page 283].

Exercise 8.16. Suppose \( \{X_k\} \) are independent identically distributed and integrable. Show that \( \frac{1}{n} S_n \) is uniformly integrable, so \( \frac{1}{n} S_n \to m \) in \( L_1 \). That is, \( E|\frac{1}{n} S_n - m| \to 0 \).

\(^5\)Etemadi’s proof allows \( \delta = 0 \) by using geometric subsequences!
8. Sums of independent random variables

**Hint:** Without loss of generality we can assume that mean is zero. Use symmetry: \( \int_{|S_n|>an} X_k dP = \int_{|S_n|>an} X_1 dP \) for \( k \leq n \) to show that \( \int_{|S_n|>an} \frac{1}{n}|S_n| dP \to 0 \) as \( n \to \infty \). Deduce from this that

\[
\lim_{a \to \infty} \sup_n \int_{|\frac{1}{n}S_n|>a} \frac{1}{n}|S_n| dP = 0
\]

**Exercise 8.17.** Let \( X_1, X_2, \ldots \) be independent random variables. Show that \( A = \{ \omega : \frac{1}{n}(X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)) \to 17 \} \) is a tail event.

**Exercise 8.18.** Let \( X_1, X_2, \ldots \) be independent random variables. Show that

\[
A = \{ \omega : \sum_{n=1}^{\infty} X_n(\omega) \text{ converges} \}
\]

is a tail event.
Chapter 9

Weak convergence

This is based on [Billingsley, Section 25]

1. Convergence in distribution

A cumulative distribution function $F$ can have at most a countable number of discontinuity points. In fact, the set $\{x : F(x) - F(x^-) \geq 1/n\}$ can have at most $n$ points. (This observation will be used in many proofs.)

**Definition 9.1.** Let $F_n$, $F$ be cumulative distribution functions. We say that $F_n \xrightarrow{D} F$ if $F_n(x) \to F(x)$ for every point $x$ of continuity of $F$.

We say that $X_n \xrightarrow{D} X$ if $F_n \xrightarrow{D} F$.

We first show how to use the definition to prove weak convergence.

**Example 9.1.** Suppose $\{X_k\}$ are independent exponential. Then $\max_{1 \leq k \leq n} X_k - \ln n \xrightarrow{D} Y$ where $Y$ has the Gumbel distribution: $P(Y \leq x) = \exp(-e^{-x})$.

Indeed, $P(\max_{1 \leq k \leq n} X_k - \ln n \leq x) = P(\max_{1 \leq k \leq n} X_k \leq x + \ln n) = P(X_1 \leq x + \ln n)^n = (1 - e^{-x \ln n})^n = (1 - e^{-x})^n \to e^{-e^{-x}}$

The following example illustrates that we cannot require convergence for all $x \in \mathbb{R}$.

**Example 9.2.** Suppose $X_n$ are uniform $U(0, 1/n)$ with

$$F_n(x) = \begin{cases} 1 & x > 1/n \\ nx & x \in [0, 1/n] \\ 0 & x < 0 \end{cases}$$

It is clear that $X_n \to 0$ with probability one, so we expect (and can prove, see Theorem 9.2 below) that $X_n \xrightarrow{D} 0$, i.e. that $F_n(x) \to F(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$. And indeed, $F_n(x) = 0$ for $x < 0$ and $F_n(x) \to 1$ for $x > 0$. But note that $F_n(0) = 0$ does not converge to $F(0)$. 103
The following example illustrates that a popular interpretation of weak convergence as “approximating all probabilities” for \( X_n \) by the asymptotic probabilities for \( X \) has significant restrictions.

**Example 9.3.** Suppose \( P(X_n = k) = 1/n \). Then \( \frac{1}{n} X_n \overset{D}{\to} U(0, 1) \). Indeed, \( F_n(x) = [nx+1]/n \to x \). Note however that \( P(\frac{1}{n} X_n \in V) \) may fail to converge to \( \lambda(V) \) for some Borel sets \( V \in \mathcal{B} \).

In view of Example 9.3, it is interesting to have a criterion where a stronger form of weak convergence holds.

**Theorem 9.1** (Scheffe’s theorem). Suppose \( X_n \) has a density \( f_n(x) \) with respect to a (possibly infinite, possibly discrete) measure \( \nu(dx) \) on \( \mathbb{R} \). If \( f_n(x) \to f(x) \) pointwise and \( f \) is a density of a random variable \( X \), then

\[
\sup_U |P(X_n \in U) - P(X \in U)| \to 0
\]

**Proof.** Consider \( g_n = f - f_n \). Then \( g_n^+ \to 0 \) and \( 0 \leq g_n^+ \leq f \) so by the dominated convergence theorem \( \int g_n^+ \nu(dx) \to 0 \). Now \( \int |g_n|\nu(dx) = \int_{g_n > 0} g_n \nu(dx) - \int_{g_n \leq 0} g_n \nu(dx) \). Since \( \int g_n \nu(dx) = 0 \) we have \( \int_{g_n \leq 0} g_n \nu(dx) = -\int_{g_n > 0} g_n \nu(dx) \) so

\[
\int |g_n|\nu(dx) = 2 \int g_n^+ \nu(dx) \to 0
\]

Thus \( |P(X_n \in U) - P(X \in U)| \leq \int_U |g_n(x)|\nu(dx) \to 0 \) for any \( U \in \mathcal{B} \). \( \square \)

**Example 9.4.** Suppose \( X_n \) is binomial \( \text{Bin}(n, p = \lambda/n) \). Then \( X_n \overset{D}{\to} Y \) where \( Y \) is Poiss(\( \lambda \)). Indeed, the density with respect to the counting measure \( \nu \) converges pointwise

\[
P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1) \ldots (n-k+1) \lambda^k}{k!} (1-\lambda/n)^n \to e^{-\lambda} \lambda^k / k!
\]

Thus in this case \( P(X_n \in A) \to P(Y \in A) \) for all \( A \). For example,

\[
P(X_n \text{ is even}) \to e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2k!}
\]

Similarly, as the number of degrees of freedom \( d \to \infty \), the density of student \( T_d \) distribution converges to the standard normal density.

**Theorem 9.2.** If \( X_n \overset{P}{\to} X \) then \( X_n \overset{D}{\to} X \).

**Proof.** Let \( x \) be a point of continuity of \( F(x) \). Then

\[
P(X_n \leq x) = P(X_n \leq x, |X_n - X| \leq \varepsilon) + P(X_n \leq x, |X_n - X| > \varepsilon)
\]

So

\[
P(X_n \leq x) \leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon)
\]

Similarly,

\[
P(X \leq x - \varepsilon) \leq P(X_n \leq x) + P(|X_n - X| > \varepsilon)
\]

So

\[
F(x - \varepsilon) \leq \lim \inf P(X_n \leq x) \leq \lim \sup P(X_n \leq x) \leq F(x + \varepsilon)
\]

Taking the limit \( \varepsilon \to 0 \),

\[
F(x^-) \leq \lim \inf P(X_n \leq x) \leq \lim \sup P(X_n \leq x) \leq F(x)
\]
Remark 9.3. If \( X_n \xrightarrow{D} a \) for a deterministic random variable \( a \) then \( X_n \xrightarrow{P} a \). Indeed, \( P(X_n \leq a - \varepsilon) \to 0 \) and \( P(X_n \leq a + \varepsilon) \to 1 \) so
\[
P(|X_n - a| > \varepsilon) \leq P(X_n \leq a - \varepsilon) + 1 - P(X_n \leq a + \varepsilon) \to 0
\]

Theorem 9.4 (Slutsky’s Theorem). Suppose \((X_n, Y_n)\) are defined on the same probability space. If \( X_n \xrightarrow{D} X \) and \( Y_n \xrightarrow{P} 0 \) then \( X_n + Y_n \xrightarrow{D} X \)

Proof. Take \( y' < y'' \) two continuity points of the law of \( X \) and \( y' < x - \varepsilon < x < x + \varepsilon < y'' \). Then
\[
P(X_n \leq y') - P(|Y_n| > \varepsilon) \leq P(X_n + Y_n \leq x) \leq P(X_n \leq y'') + P(|Y_n| > \varepsilon)
\]
So
\[
F(y') \leq \lim \inf P(X_n + Y_n \leq x) \leq \lim \sup P(X_n + Y_n \leq x) \leq F(y'')
\]
We now note that the set \( \{ x : F(x-) - F(x) \geq 1/n \} \) has at most \( n \) points, so the set of all discontinuities of \( F \) is at most countable. Therefore, if \( x \) is a continuity point of \( F \) we can find continuity points \( y' < x < y'' \) that are arbitrarily close to \( x \). Thus taking a sequence \( y' \to x \) and \( y'' \to x \) of such points we get
\[
F(x) \leq \lim \inf P(X_n + Y_n \leq x) \leq \lim \sup P(X_n + Y_n \leq x) \leq F(x)
\]

The following corollary is often useful. (Its proof requires tightness!).

Corollary 9.5. Suppose \((X_n, Y_n)\) are defined on the same probability space. If \( X_n \xrightarrow{D} X \) and \( Y_n \xrightarrow{P} c \) then \( X_n Y_n \xrightarrow{D} cX \).

Proof. See Exercise 9.2

2. Fundamental results

Theorem 9.6 (Skorohod’s theorem). Suppose \( X_n \xrightarrow{D} X \) i.e. \( F_n \xrightarrow{D} F \). Then there exist a probability space \((\Omega, \mathcal{F}, P)\) and random variables \( Y_n, Y \) with CDF \( F_n, F \) on \((\Omega, \mathcal{F}, P)\) such that \( Y_n \to Y \) for all \( \omega \in \Omega \).

Proof. We choose \( \Omega = (0, 1) \) with Lebesgue measure. Recalling the quantile function (4.3) we define
\[
Y_n(\omega) = \inf \{ x : F_n(x) \geq \omega \}
\]
\[
Y(\omega) = \inf \{ x : F(x) \geq \omega \}
\]
Recall that \( Y(\omega) \leq x \) iff \( F(x) \geq \omega \) so \( Y(\omega) > x \) implies \( F(x) < \omega \).

Given \( \varepsilon > 0 \) choose \( Y(\omega) - \varepsilon < x < Y(\omega) \) such that \( F(x^-) = F(x) \).

Since \( F_n(x) \to F(x) \), this implies that \( F_n(x) < \omega \) for large \( n \). Thus \( Y_n(\omega) > x > Y(\omega) - \varepsilon \).
Since \( \varepsilon > 0 \) this shows that \( \lim \inf Y_n(\omega) \geq Y \).

Now choose \( \omega < \omega' \) and a continuity point \( y \) of \( F \) such that \( Y(\omega') < y < Y(\omega') + \varepsilon \). The first inequality then implies that \( \omega < \omega' < F(y) \), so for large \( n \) we have \( F_n(y) > \omega \). Thus \( Y_n(\omega) \leq y < Y(\omega') + \varepsilon \). This shows that \( \lim \sup Y_n(\omega) \leq Y(\omega) \) for all points \( \omega \) of continuity of \( Y \).
Note that \( Y \) is an increasing function so it can have at most countable number of discontinuities.
At such points we re-define $Y_n(\omega)$ to be $Y(\omega)$. This changes $Y_n$ on the set of measure zero, so does not affect the result.

**Theorem 9.7** (Portmanteau Theorem). The following conditions are equivalent:

(i) $X_n \xrightarrow{D} X$
(ii) $E(f(X_n)) \to E(f(X))$ for every bounded continuous function $f$
(iii) $E(f(X_n)) \to E(f(X))$ for every bounded Lipschitz (uniformly continuous) function $f$
(iv) $P(X_n \in U) \to P(X \in U)$ for every Borel set $U$ such that $P(X \in \delta U) = 0$

**Proof.** (Omitted in 2019) (1)⇒(2) Using Theorem 9.6 we have $f(Y_n) \to f(Y)$ so by Lebesgue’s dominated convergence theorem (Theorem 6.9), the integrals converge. Note that this proof works for $\mathbb{R}$ but not for $\mathbb{R}^2$, so it is of interest to have a direct proof that will not use Theorem 9.6. See [Billingsley, Theorem 29.1]

(2)⇒(3) is obvious

(3)⇒(1) Fix a point of continuity $x_0$ of $F$ and let

$$f(x) = \begin{cases} 1 & x \leq x_0 \\ \text{linear} & x_0 < x < x_0 + \varepsilon \\ 0 & x > x_0 + \varepsilon \end{cases}$$

Then $F_n(x_0) \leq E(f(X_n)) \to Ef(X) \leq F(x_0 + \varepsilon)$ so $\limsup_n F_n(x_0) \leq F(x_0)$.

Next, take

$$f(x) = \begin{cases} 1 & x \leq x_0 - \varepsilon \\ \text{linear} & x_0 - \varepsilon < x < x_0 \\ 0 & x \geq x_0 \end{cases}$$

Then $F_n(x_0) \geq E(f(X_n)) \toEf(X) \geq F(x_0 - \varepsilon)$ Thus $\liminf_n F_n(x_0) \geq F(x_0)$.

(1)⇒(2) [Second proof] Suppose $f$ is continuously differentiable and $f' = 0$ outside of a finite interval $[-K, K]$. Then from Fubini’s theorem\(^1\) we get

$$Ef(X_n) = f(0) + \int_0^K f'(t) P(X_n > t) dt - \int_{-K}^0 f'(t) P(X_n \leq t) dt$$

Since $P(X_n \leq t) \to P(X \leq t)$ except for a countable (Lebesgue-measure zero) set of $t$, by Lebesgue dominated convergence theorem we get $Ef(X_n) \to Ef(X)$.

\(^1\) $f(x) = f(0) + \int_{x>0} f'(t) dt - \int_{x<0} f'(t) dt$
Theorem 9.8 (Continuous Mapping Theorem). As an immediate corollary, we get an important result.

Example 9.5. In the setting of Example 9.3, we have \( f \) such that by using a diagonal method, there is a subsequence and a subsequence \( \{q_k\} \) such that \( |g(x) - f(x)| < \varepsilon \) for \( x \in [-K, K] \) and \( g \) is bounded by 2\( M \). Finally, extend \( g \) to \( \mathbb{R} \) in such a way that \( g(x) = 0 \) outside of \([-K - 1, K + 1] \) and without increasing the bound \( |g(x)| \leq 2M \). Then

\[
|Ef(X) - Ef(X)| \leq 2\varepsilon + \int_{|x| \leq K, |x| \leq K} (f(X_n) - f(X)) \, dP \\
\leq 4\varepsilon + \int_{|x| \leq K, |x| \leq K} (g(X_n) - g(X)) \, dP \\
\leq 6\varepsilon + |Ef(X_n) - Ef(X)|
\]

Since \( |Ef(X_n) - Ef(X)| \to 0 \) by the previous part of the proof, and \( \varepsilon > 0 \) is arbitrary, we see that \( Ef(X_n) \to Ef(X) \).

\[
\square
\]

As an immediate corollary, we get an important result.

Theorem 9.9 (Helly, Prokhorov). If \( X_n \xrightarrow{D} X \) and \( f \) is a continuous (but perhaps unbounded) function then \( f(X_n) \xrightarrow{D} f(X) \).

Example 9.9. In the setting of Example 9.9, we have \( Ef(X_n/n) = \frac{1}{n} \sum_{k=1}^{n} f(k/n) \to \int_{0}^{1} f(x) \, dx \)

Definition 9.2. A sequence of probability measures \( \mu_n \) on \( \mathbb{R} \) is tight if for every \( \varepsilon > 0 \) there exists \( K \) such that \( \mu_n([-K, K]) > 1 - \varepsilon \).

It is clear that if \( X_n \xrightarrow{D} X \) then \( X_n \) is tight.

Theorem 9.10. If \( \mu_n \) is a tight family of probability measures then there is a probability measure \( \mu \) and a subsequence \( n_k \to \infty \) such that \( \mu_{n_k} \xrightarrow{D} \mu \).

Proof. Since \( F_n(r) \) is a bounded sequence of numbers, there is a subsequence that converges. In fact, by using a diagonal method, there is a subsequence \( n_k \) such that \( F_{n_k}(r) \to G(r) \) for all \( r \in \mathbb{Q} \).

To see this, enumerate all rational numbers \( q_1, q_2, \ldots \). Since \([0, 1]\) is compact, we can choose a sequence \( n(k) = n_1(k) \) such that \( F_{n(k)}(q_1) \) converges to, say, \( G(q_1) \). Choose a subsequence \( n_2(k) \) of \( n_1(k) \) such that \( F_{n_2(k)}(q_1) \) converges to, say, \( G(q_1) \) and so on.

\[
\begin{array}{ccccccccc}
n_1(1) & n_1(2) & n_1(3) & \ldots & \ldots & F_{n_1(k)}(q_1) & \to G(q_1) \\
n_2(1) & n_2(2) & n_2(3) & \ldots & \ldots & F_{n_2(k)}(q_2) & \to G(q_2) \\
n_3(1) & n_3(2) & n_3(3) & \ldots & \ldots & F_{n_3(k)}(q_3) & \to G(q_3) \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
n_j(1) & n_j(2) & n_j(3) & \ldots & n_j(j) & \ldots & F_{n_j(k)}(q_j) & \to G(q_j) \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
\end{array}
\]

Then the diagonal subsequence \( m_k := n_k(k) \) has the property that \( F_{m_k}(q) \to G(q) \) for every \( q \in \mathbb{Q} \).

Define

\[
F(x) = \inf \{ G(r) : r > x \}
\]
Note that \( F(x) = \lim_{r \downarrow x} G(r) \), so \( F \) is non-decreasing and right-continuous. By tightness, \( F(x) < \varepsilon \) if \( x < K \) and \( F(x) > 1 - \varepsilon \) if \( x > K \). Next we check that \( F \) is right-continuous:

Now we verify the weak convergence. Let \( x \) be a point of continuity of \( F \). Choose \( r_k \uparrow x \) and \( r'_k \downarrow x \).

Then

\[
F_n(r_k) \leq F_n(x) \leq F_n(r'_k)
\]

so for every \( k \) we have

\[
G(r_k) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq G(r'_k)
\]

But \( G(r'_k) \rightarrow F(x) \) as \( k \rightarrow \infty \). And for any \( \varepsilon_k > 0 \) converging to 0 we have \( G(r_k) \geq F(r_k - \varepsilon_k) \rightarrow F(x) \) by continuity.

□

**Example 9.6.** Suppose \( X_n \) are uniform on \((0,n)\). Then

\[
F_n(x) = \begin{cases} 
0 & x < 0 \\
x/n & 0 \leq x \leq n \\
1 & x > n 
\end{cases}
\]

So \( F_n(x) \rightarrow F_\infty(x) \) for all \( x \). Clearly \( F_\infty(x) \) is not a cumulative distribution function, and \( X_n \) is not a tight sequence.

We will need the following corollary.

**Theorem 9.10.** If \( \mu_n \) is a tight family of probability measures and if each subsequence converges to the same probability measure \( \mu \) then \( \mu_n \overset{D}{\rightarrow} \mu \).

**Proof.** Suppose \( \mu_n \) fails to converge to \( \mu \) with CDF \( F \). Then there is a point of continuity \( x \) of \( F \) and an infinite sequence \( n_k \) such that \( |F_{n_k}(x) - F(x)| > \delta \) for all \( k \). Since subsequence \( \mu_{n_k} \) is tight, choose a convergent subsequence. By assumption, this sequence converges to \( \mu \), so \( F_{n_{k_j}}(x) \rightarrow F(x) \), which contradicts that \( |F_{n_k}(x) - F(x)| > \delta \) for all \( k \). □

Recall Definition 6.3: Family \( \{X_n\} \) is uniformly integrable if for every \( \varepsilon > 0 \) there is \( K \) such that

\[
\int_{|X_n| > K} X_n^2 dP < \varepsilon.
\]

**Proposition 9.11.** If \( \{X_n\} \) is uniformly integrable, then \( \sup_n E|X_n| < \infty \)

**Proof.** (This should have been an assigned exercise!) \( E|X_n| = \int_{|X_n| \leq K} |X_n| dP + \int_{|X_n| > K} |X_n| dP < K_\varepsilon + \varepsilon \). □

**Theorem 9.12.** Suppose \( X_n \overset{D}{\rightarrow} X \) and \( \{X_n\} \) is uniformly integrable. Then \( X \) is integrable and \( E(X_n) \rightarrow E(X) \).

**First proof.** From Theorem 9.6 there exists a sequence \( Y_n \rightarrow Y \) such that \( E(Y_n) = E(X_n) \). [For this proof, we need to know that \( |Y_n - Y| \) is uniformly integrable!] By Lebesgue’s dominated convergence theorem (See Remark 6.13), \( E(Y_n) \rightarrow E(Y) \). □
Second proof. The first step is to prove that $X$ is integrable, which we will omit.\footnote{Choose bounded continuous $f_K$ as in the main part of the proof but apply it to $|X_n|$ so that $0 \leq f_K(|X_n|) \leq |X_n|$. Then $E(|X|I_{|X| \leq K}) \leq E f_K(|X|)$. But $Ef_K(|X|) = \lim_{n \to \infty} Ef_K(|X_n|)$. And $Ef_K(|X_n|) \leq E|X_n| \leq M$ by Proposition 9.11. So $E|X| = \lim_{K \to \infty} E(|X|I_{|X| \leq K}) \leq M < \infty.$}

Given $\varepsilon > 0$ choose $K$ such that $\int_{|X_n| > K} |X_n|dP < \varepsilon$. Since $X$ is integrable, we can increase $K$ to ensure that we also have $\int_{|X| > K} |X|dP < \varepsilon$.

Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise-linear bounded continuous function such that $f(x) = x$ for $x \in [-K, K]$ and $f(x) = 0$ for $x \not\in [-K - 1, K + 1]$. (Draw the graph. Note that $f = f_K$ depends on $K$.) By Theorem 9.7, $E(f(X_n)) \to E(f(X))$. On the other hand, $X_n = f(X_n)$ for $|X_n| \leq K$ and $|f(x)| \leq |x|$ for all $x$, so

\[
|E(X_n) - E(X)| \leq |E(X_n) - E(f(X_n))| + |E(X) - E(f(X))| + |Ef(X_n) - Ef(X)|
\]

\[
\leq 2 \int_{|X_n| \geq K} |X_n|dP + 2 \int_{|X| \geq K} |X|dP + |Ef(X_n) - Ef(X)|
\]

\[
\leq 4\varepsilon + |Ef(X_n) - Ef(X)|
\]

Since $|Ef(X_n) - Ef(X)| \to 0$ this implies convergence.

The proof of the next result contains solution of (a generalization of) Exercise 6.6.

**Corollary 9.13.** Suppose $\sup_n E|X_n|^{r+\delta} < \infty$ for a natural $r$ and $\delta > 0$. If $X_n \overset{D}{\to} X$ then $E(|X|^r) < \infty$ and $E(X_n^r) \to E(X^r)$.

**Proof.** We verify that $|X_n|^r$ is uniformly integrable, compare Exercise 6.6.

\[
\int_{|X_n| > t} |X_n|^r dP = \int_{|X_n| > t} |X_n|^r 1_{dP} \leq \int_{|X_n| > t} |X_n|^r \frac{|X_n|^\delta}{t^\delta} dP \leq \frac{1}{t^\delta} \sup_n E|X_n|^{r+\delta}
\]

\[\square\]

---

**Required Exercises**

**Exercise 9.1.** Suppose $\{X_k\}$ are independent uniform $U(0,1)$ random variables. Show that

\[
n \min_{1 \leq k \leq n} X_k \overset{D}{\to} Y
\]

and determine the law of $Y$.

**Exercise 9.2.** Prove Corollary 9.5: if $X_n \overset{P}{\to} c$ for a constant $c$ and $Y_n \overset{D}{\to} Y$, show that $X_nY_n \overset{D}{\to} cY$.

**Exercise 9.3.** Suppose $X_n \overset{D}{\to} X$. Show that the laws of $X_n$ are tight.

**Exercise 9.4.** Suppose $X_n \in \mathbb{Z}$ and $X_n \overset{D}{\to} X$. Show that $P(X \in \mathbb{Z}) = 1$ and that $P(X_n = k) \to P(X = k)$ for every $k \in \mathbb{Z}$.
Exercise 9.5. Suppose $X_n$ has density $f_n(x) = 1 + \cos(2\pi nx)$ on $[0,1]$. Prove that $X_n \xrightarrow{D} X$ (and determine the law of $X$).

Exercise 9.6. Suppose $E(X_n^2) = 1$. Show that $F_n$ is tight.

Exercise 9.7. Suppose $E(X_n^2) = 1$. Show that $\{X_n\}$ is uniformly integrable.

Exercise 9.8. Show that $X$ is integrable if and only if for every $\varepsilon > 0$ there exists $K$ such that $
int_{|X| > K} |X|dP < \varepsilon$. \hspace{1cm} (This is Corollary 6.12 on page 73)

Exercise 9.9. Suppose that $\sup_n E(|X_n|f(|X_n|)) < \infty$ for some non-decreasing function $f$ such that $\lim_{x \to \infty} f(x) = \infty$. Show that $\{X_n\}$ is uniformly integrable.

Exercise 9.10. The Lévy distance between two probability measures on $\mathbb{R}$ is defines as 
$$d(F,G) = \inf\{\varepsilon > 0 : G(x-\varepsilon) - \varepsilon \leq F(x) \leq G(x+\varepsilon) + \varepsilon \text{ for all } x\}$$
(i) Verify that this is a metric\(^3\).
(ii) Verify that $F_n \xrightarrow{D} F$ iff $d(F_n, F) \to 0$
(iii) Verify that for every probability measure $\mu$ on Borel sets of $\mathbb{R}$ there exist probability measures $\mu_n$ with finite support such that $\mu_n \xrightarrow{D} \mu$. Show further that the support can be taken from $\mathbb{Q}$, so that the space of distribution functions is separable in the Lévy metric.

Definition 9.3. We say that $(X_n, Y_n) \xrightarrow{D} (X, Y)$ if for every bounded continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ we have $E(f(X_n, Y_n)) \to E(f(X, Y))$.

Exercise 9.11. Suppose $(X_n, Y_n)$ are independent and $X_n \xrightarrow{D} X, Y_n \xrightarrow{D} Y$. Prove that $(X_n, Y_n) \xrightarrow{D} \mu$ where $\mu = F_X \otimes F_Y$ is the product measure.

Exercise 9.12. Suppose $(X_n, Y_n) \xrightarrow{D} (X, Y)$. Prove that $X_n^2 + Y_n^2$ converges in distribution.

---

\(^3\)Non-negative, triangle inequality, and $d(F, G) = 0$ only for $F = G$
Chapter 10

Characteristic functions

This is based on [Billingsley, Section 26]

1. Complex numbers, Taylor polynomials, etc

**Theorem 10.1** (Taylor polynomials). For any "smooth enough" function \( f \) we have the following identity

\[
(10.1) \quad f(x) = f(0) + f'(0)x + \cdots + \frac{f^{(n)}(0)}{n!} x^n + \frac{1}{n!} \int_0^x f^{(n+1)}(s)(x-s)^n ds
\]

**Proof.** This is integration by parts formula: Case \( n = 0 \) is

\[
\int_0^x f(s) ds = f(0) + \int_0^x f'(s) ds
\]

Suppose the formula holds for some \( n \geq 0 \). Then

\[
\int_0^x f^{(n+1)}(t)(x-s)^n ds = \frac{1}{n+1} \int_0^x f^{(n+1)}(s)(-s)^n ds = \frac{1}{n+1} \left( \int_0^x f^{(n+2)}(s)(x-s)^{n+1} ds \right)
\]

Putting this back into (10.1), we get the same formula for \( n + 1 \). \( \square \)

When \( \frac{1}{n!} \int_0^x f^{(n+1)}(s)(x-s)^n ds \to 0 \) as \( n \to \infty \) we get the series expansion for \( f(x) \). Special cases of interest in this course are:

\[
(10.2) \quad e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}
\]
10. Characteristic functions

\[ \cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \]

\[ \sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \]

We will also need sharp error estimates!!!

**Exercise 10.1.** Use Theorem 10.1 to prove (10.2).

1.1. Complex numbers. A complex number is an expression \( z = x + iy \) where \( i^2 = -1 \). Note: \( x \) is called the real part of \( z \) and \( y \) is called the imaginary part of \( z \).

The modulus of a complex number is \( |z| = \sqrt{x^2 + y^2} \). Noting that \( |z|^2 = z\bar{z} \) with complex conjugate \( \bar{z} = x - iy \), we get

\[ |z_1z_2| = |z_1||z_2| \]

Since \( |z| \) is a distance in \( \mathbb{R}^2 \), we get the triangle inequality \( |z_1 + z_2| \leq |z_1| + |z_2| \).

Arithmetic works as usual. For example, \((1 + i)^2 = 2i\). Some powers of \( i \): \( i^3 = -i, i^4 = 1, i^5 = i, i^6 = -1 \).

(Omitted in 2019) One simple explanation why arithmetics works comes from matrix interpretation: To a complex number \( a + ib \) associate the matrix \( \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \). Then 1 corresponds to \( I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \) and \( i \) corresponds to \( J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \). Under this model, multiplication of complex numbers \((a + ib)(c + id)\) corresponds to multiplication of matrices. For example, \( J^2 = -I \).

It is clear that

\[ i^n = \begin{cases} 
  i \\
  -i \\
  1 \\
  -1
\end{cases} \]

More specifically:

\[ i^n = \begin{cases} 
  (-1)^k & n = 2k \\
  (-1)^k i & n = 2k + 1
\end{cases} \]

The following is called de Moivre formula

\[ (10.3) \]

\[ e^{ix} = \cos x + i \sin x \]

**Proof.** We use (10.2):

\[ e^{ix} = \sum_{n=0}^{\infty} \frac{i^n x^n}{n!} = \sum_{k=0}^{\infty} \frac{i^{2k} x^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{i^{2k+1} x^{2k+1}}{(2k+1)!} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} + i \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = \cos x + i \sin x \]

\[ \square \]

In particular, \( e^{i(x+y)} = e^{ix} e^{iy} \) which is a just a version of the formula for \( \cos(x+y) \) and \( \sin(x+y) \).

**Exercise 10.2.** Find all complex numbers \( z \) with the property that \( z^2 = i \).
1. Complex numbers, Taylor polynomials, etc

1.2. Complex version of Taylor’s formula. Integration by parts works also for complex functions, so (10.1) holds also for $f(x) = e^{ix}$. This gives

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!} \int_{0}^{x} (x-s)^n e^{is} ds$$

Note that for $x > 0$ we have $|\int_{0}^{x} (x-s)^n e^{is} ds| \leq \int_{0}^{x} (x-s)^n ds = x^{n+1}/(n+1)$. Similarly, for $x < 0$ we have $|\int_{0}^{x} (x-s)^n e^{is} ds| \leq \int_{0}^{x} (s-x)^n = (-x)^{n+1}/(n+1)$. This gives

$$|e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}| \leq \frac{|x|^{n+1}}{(n+1)!}.$$  

However, this bound is not as good as we need for large $|x|$.

**Lemma 10.2.** For $n = 0, 1, \ldots$ we have

$$|e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}| \leq \min \left\{ \frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!} \right\}$$

**Proof.** These improved bound is based on the identity

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{(n-1)!} \int_{0}^{x} (x-s)^n (e^{is} - 1) ds$$

which for $n \geq 1$ comes from integration by parts backwards:

$$\int_{0}^{x} (x-s)^n e^{is} ds = \frac{1}{i} (x-s)^n e^{is}|_{s=0}^{x} + \frac{n}{i} \int_{0}^{x} (x-s)^{n-1} e^{is} ds = \frac{1}{i} x^n + \frac{n}{i} \int_{0}^{x} (x-s)^{n-1} e^{is} ds$$

$$= -\frac{n}{i} \int_{0}^{x} (x-s)^{n-1} ds + \frac{n}{i} \int_{0}^{x} (x-s)^{n-1} e^{is} ds$$

The error estimate is $\left| \frac{i^n}{(n-1)!} \int_{0}^{x} (x-s)^{n-1} (e^{is} - 1) ds \right| \leq 2|x|^n/n$. This gives

$$|e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}| \leq \frac{2|x|^n}{n!}.$$ 

Combining this with the previous estimate (10.5) we get (10.6). (The case $n = 0$ is triangle inequality, $|e^{ix} - 1| \leq 2$.)

We will only need (10.6) for $n = 1, 2, 3$.

$$|e^{ix} - 1| \leq \min \{|x|, 2\}$$

$$|e^{ix} - (1 + ix)| \leq \min \left\{ \frac{1}{2} x^2, 2|x| \right\}$$

$$|e^{ix} - (1 + ix - \frac{1}{2} x^2)| \leq \min \left\{ \frac{1}{6} |x|^3, x^2 \right\}$$

Since we need only these three expansion, it is simpler to give a direct proof.
Proof. An alternative proof of the first inequality:

\[ e^{-ix} - 1 = \int_0^x (e^{is})' ds = i \int_0^x e^{is} ds, \]
so (assuming \( x > 0 \)) we get \(|e^{ix} - 1| \leq \int_0^x 1 ds = x \), and \(|e^{ix} - 1| \leq |e^{ix}| + 1 = 2 \). Same calculations can be repeated for \( x < 0 \), giving (10.7).

From the above, we get \( e^{-ix} - 1 - ix = i \int_0^x (e^{is} - 1) ds \) so \(|e^{-ix} - 1 - ix| \leq 2|x|\).

Next, we integrate by parts: \( e^{-ix} - 1 = i \int_0^x s e^{is} ds = i \int_0^x (s-x)e^{is} ds = i(s-x)e^{i\pi} - i^2 \int_0^x (s-x)e^{i\pi} ds \), so (assuming \( x > 0 \)), we get \(|e^{-ix} - 1 - ix| \leq \int t^2 (x-s) ds = x^2/2\).

Same calculations can be repeated for \( x < 0 \), giving (10.8). \( \square \)

Proof. An alternative proof of the first inequality:

\[ |e^{ix} - 1| = \sqrt{2(1 - \cos x)} \]

Clearly, \( \sqrt{2(1 - \cos x)} \leq 2 \). On the other hand, \( \cos x \geq 1 - x^2/2 \), as \( f(x) = \cos x - 1 - x^2/2 \) has \( f(0) = 0 \), \( f'(x) = x - \sin x > 0 \) for \( x > 0 \). So \( \sqrt{2(1 - \cos x)} \leq |x| \) \( \square \)

1.3. Integrating complex-valued random variables. If \( Z : \Omega \to \mathbb{C} \) is a random variable, then \( Z(\omega) = X(\omega) + iY(\omega) \). The property we will need is integration of products of complex-valued expressions in independent random variables:

Proposition 10.3. If \( Z_1 = X_1 + iY_1 \) and \( Z_2 = X_2 + iY_2 \) are independent then \( E(Z_1 Z_2) = E(Z_1) E(Z_2) \).

Proof. Write \( Z_1 Z_2 = X_1 X_2 - Y_1 Y_2 + i(X_1 Y_2 + Y_1 X_2) \) and integrate each term. \( \square \)

Complex version of Theorem 6.17(ii) is a bit more difficult.

Proposition 10.4. If \( |Z| \) is integrable, then \( Z \) is integrable and \( |E(Z)| \leq E(|Z|) \)

Proof. Writing \( Z = X + iY \), we have \(|X| \leq |Z| \) and \(|Y| \leq |Z| \) so \( X, Y \) are integrable. The inequality says that \( \sqrt{(EX)^2 + (EY)^2} \leq E \sqrt{X^2 + Y^2} \). This is Jensen’s inequality for the convex function \( d(x, y) = \sqrt{x^2 + y^2} \) (the distance in \( \mathbb{R}^2 \) is a convex function!). \( \square \)

2. Characteristic functions

Definition 10.1. The characteristic function of a real-valued random variable \( X \) is

\[ \varphi(t) = E e^{itX} \]

In principle, \( \varphi(t) \) contains the same amount of “information” as a pair of functions \( E \cos(tX) \) and \( E \sin(tX) \). But it is convenient to use the standard properties of the exponential function.
Remak 10.5. Of course, $\varphi(0) = 1$ and $|\varphi(t)| \leq 1$ for all $t$. In fact, $\varphi(t)$ is uniformly continuous:

$$|\varphi(t + h) - \varphi(t)| \leq \int_{\mathbb{R}} |e^{ihx} - 1| F(dx) \to 0 \text{ as } h \to 0$$

This is a consequence of the Lebesgue dominated convergence theorem, but it is a good exercise to deduce it directly from (10.7):

$$\int_{\mathbb{R}} |e^{ihx} - 1| F(dx) = \int_{|x| > 1/\sqrt{n}} |e^{ihx} - 1| F(dx) + \int_{|x| \leq 1/\sqrt{n}} |e^{ihx} - 1| F(dx)$$

$$\leq \int_{|x| > 1/\sqrt{n}} 2F(dx) + \int_{|x| \leq 1/\sqrt{n}} \sqrt{n}F(dx) \leq 2P(|X| > 1/\sqrt{n}) + \sqrt{n} \to 0 \text{ as } h \to 0$$

The characteristic function $\varphi(t)$ can be thought as the moment generating function $M(it)$ applied to complex argument $it$.

Example 10.1. The moment generating function of the exponential random variable with density $f(x) = e^{-x}1_{x > 0}$ is defined only for $t < 1$ and is given by $M(t) = \frac{1}{1-t}$. The characteristic function $\varphi(t) = \frac{1}{1-it}$ is defined for all real $t$.

Example 10.2. The moment generating function of the standard normal random variable with density $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ is given by $M(t) = e^{t^2/2}$. The characteristic function is $\varphi(t) = e^{-t^2/2}$.

The basic properties to establish are:

- The characteristic function uniquely determines the distribution.
- $X_n \overset{D}{\to} X$ iff $\varphi_n(t) \to \varphi(t)$ for all $t$
- If $X, Y$ are independent then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

The last property is the easiest, so we can prove it now:

Proposition 10.6. If $X, Y$ are independent and $S = aX + bY + c$ then $\varphi_S(t) = \varphi_X(at)\varphi_Y(bt)e^{ict}$

Proof. Use algebra $e^{itS} = e^{iatX}e^{ibtY}e^{ict}$ and Proposition 10.3. \quad \square

2.1. Useful estimates. We conclude with some useful estimates that follow directly from (10.6)

If $\varphi(t) = Ee^{itX}$ and $X$ is square-integrable, then

$$|\varphi(t) - 1| \leq E(\min\{|tX|, 2\})$$

$$|\varphi(t) - (1 + itE(X))| \leq E\left(\min\left\{\frac{1}{2}(tX)^2, 2|tX|\right\}\right)$$

$$|\varphi(t) - (1 + itE(X) - \frac{t^2}{2}E(X^2))| \leq E\left(\min\left\{\frac{1}{6}|tX|^3, (tX)^2\right\}\right)$$

2.2. Moments and derivatives. Coefficients in (10.13) are in fact given by Taylor expansion for $\varphi$.

Theorem 10.7. If $E(|X|^n) < \infty$ then $\varphi(t)$ has the $n$-th derivative and $it^nEX^n = \varphi^{(n)}(0)$.

We note that the theorem has partial converse: if $\varphi(t)$ has a derivative of even order $2k$ then $E(|X|^{2k}) < \infty$
10. Characteristic functions

Proof. We verify only the property for the first two moments:
\[ \frac{\varphi(t+h)-\varphi(t)}{h} - E(iXe^{itX}) = E \left( e^{itX} \frac{e^{ihX} - 1 - ihX}{h} \right) \]
and we note that
\[ \left| \frac{e^{ihX} - 1 - ihX}{h} \right| \leq 2|X| \]
by (10.8). So we can apply the dominated convergence theorem

\[ \varphi'(t) = \lim_{h \to 0} E \left( E^{itX} \frac{e^{ihX} - 1 - ihX}{h} \right) = E \left( e^{itX} \lim_{h \to 0} \frac{e^{ihX} - 1 - ihX}{h} \right) = 0 \]

A similar argument, starting with
\[ \frac{\varphi'(t+h) - \varphi'(t)}{h} - E(i^2X^2e^{itX}) = E \left( X^2e^{itX} \frac{e^{ihX} - 1 - ihX}{h} \right) \]
gives \[ \varphi''(t) = E(i^2X^2e^{-itX}). \]

(Omitted in 2019)

Theorem 10.8. If \( F \) has a density then \( \varphi(t) \to 0 \) as \( t \to \infty \).

Proof. Suppose \( f \) is the density. Then for every \( \varepsilon > 0 \) there is a step function \( g = \sum \alpha_k I_{(a_k,b_k]} \) such that \( \int |f - g|dx < \varepsilon \), so \( |\varphi(t) - \int e^{itx}g(x)dx| < \varepsilon \) for all \( t \). Now \( \int e^{itx}g(x)dx = \sum \alpha_k \frac{e^{itb_k} - e^{ita_k}}{it} \to 0 \) as \( t \to \infty \).

3. Uniqueness

The fact that characteristic function determines distribution uniquely can be proved in many ways. For example, one can use Weierstrass theorem — for a sketch of such a proof, see [Billingsley, Exercise 26.19]. Or one can use convolutions – such a proof appears e.g. in [Resnik]. We will follow [Billingsley].

3.1. Inversion formula.

Theorem 10.9 (Inversion Formula). If a cumulative distribution function \( F \) has characteristic function \( \varphi(t) \) then for points of continuity of \( F \) we have

\[ F(b) - F(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-ita} - e^{-itb} \varphi(t)dt \]

Since points of continuity are dense, we have uniqueness.

Corollary 10.10 (uniqueness). If \( F,G \) have the same characteristic function \( \varphi(t) \) then \( F(x) = G(x) \) for all \( x \).

Proof. Denote by \( I_T \) the right hand side of (10.14). Fubini’s theorem gives

\[ I_T = \frac{1}{2\pi} \int_{\mathbb{R}} \left( \int_{-T}^{T} e^{it(x-a)} - e^{it(x-b)} \frac{dt}{it} \right) F(dx) \]
(To use (7.3) notice that the integrand is bounded!) We then re-write the inner integral using the fact that \(\sin\) is odd while \(\cos\) is even:

\[
\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \, dt = \frac{1}{\pi} \int_{0}^{T} \frac{\sin(t(x-a)) - \sin(t(x-b))}{t} \, dt
\]

We need the following non-obvious fact (see Example 7.2 on page 85)

\[
\lim_{T \to \infty} \int_{0}^{T} \frac{\sin t}{t} \, dt = \frac{\pi}{2}.
\]

Noting that

\[
\lim_{T \to \infty} \int_{0}^{T} \frac{\sin tx}{t} \, dt = \lim_{T \to \infty} \int_{0}^{Tx} \frac{\sin u}{u} \, du = \begin{cases} 
\frac{\pi}{2} & \text{if } x > 0 \\
0 & \text{if } x = 0 \\
-\frac{\pi}{2} & \text{if } x < 0
\end{cases}
\]

we get

\[
\lim_{T \to \infty} \frac{1}{\pi} \int_{0}^{T} \frac{\sin(t(x-a))}{t} \, dt - \frac{1}{\pi} \int_{0}^{T} \frac{\sin(t(x-b))}{t} \, dt = \begin{cases} 
-\frac{1}{2} + \frac{1}{2} = 0 & \text{if } x < a \\
0 + \frac{1}{2} = \frac{1}{2} & \text{if } x = a \\
\frac{1}{2} + \frac{1}{2} = 1 & \text{if } a < x < b \\
\frac{1}{2} - 0 = \frac{1}{2} & \text{if } x = b \\
\frac{1}{2} - \frac{1}{2} = 0 & \text{if } x > b
\end{cases}
\]

\(\square\)

The following application of inversion formula deals with the densities.

**Theorem 10.11.** If \(\varphi(t)\) is integrable then \(F\) has density

\[(10.15)\quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) \, dt\]

**Proof.** Taking \(T \to \infty\) in (10.14) we get

\[
\frac{F(x+h) - F(x)}{h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} - e^{-it(x+h)} \varphi(t) \, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} 1 - e^{-ith} e^{-itx} \varphi(t) \, dt
\]

and the formula for \(f(x) = F'(x)\) follows from Lebesgue’s dominated convergence theorem, as \(\lim_{h \to 0} 1 - e^{-ith} = 1\).

\(\square\)

As an application of Theorem 10.11 we deduce the following.

**Proposition 10.12.** The characteristic function of the Cauchy distribution with \(F(x) = \frac{1}{2} + \arctan(x)/\pi\) is \(\varphi(t) = e^{-|t|}\)

**Proof.** Since \(\varphi\) is integrable, we compute the density of the Cauchy distribution:

\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-|t|} \, dt = \frac{1}{2\pi} \int_{-\infty}^{0} e^{-itx} e^{t} \, dt + \frac{1}{2\pi} \int_{0}^{\infty} e^{-itx} e^{-t} \, dt
\]

\[
= \frac{1}{2\pi} \left( \frac{1}{1-ix} + \frac{1}{1+ix} \right) = \frac{1}{\pi} \frac{1}{1 + x^2}
\]

\(\square\)
4. The continuity theorem

**Theorem 10.13.** Let $F_n, F$ be cumulative distribution functions with characteristic functions $\varphi_n$ and $\varphi$. Then the following conditions are equivalent:

(i) $F_n \xrightarrow{D} F$

(ii) $\varphi_n(t) \to \varphi(t)$ for each $t$.

Note that $\varphi_n(t)$ may converge without $F_n \xrightarrow{D} F$, see Exercise 10.6.

**Proof.** If $X_n \xrightarrow{D} X$ then $\varphi_n(t) \xrightarrow{D} \varphi(t)$ for all $t$ by Portmanteau Theorem 9.7.

The difficult part of proof is to show that the convergence of $\varphi_n(t) \to \varphi(t)$ for each $t$ implies $F_n \xrightarrow{D} F$. The plan of proof for the converse implications is as follows:

- Show that $\varphi_n(t) \xrightarrow{D} \varphi(t)$ implies tightness.
- Use Prokhorov’s theorem (Theorem 9.9) to deduce that $F_n$ has convergent subsequences
- From the uniqueness theorem (Corollary 10.10) we verify that all limiting distributions $F$ are the same.
- By Theorem 9.10 we deduce convergence $F_n \xrightarrow{D} F$.

Clearly, the first step is the where the difficulty lies.

**Lemma 10.14.** If $\varphi_n(t) \xrightarrow{D} \varphi(t)$ for all $t$ in a neighborhood of 0 then $\{X_n\}$ is tight.

**Proof.** Since $\varphi(0) = 1$, and $\varphi$ is continuous at $t = 0$, for all $u$ small enough $\frac{1}{u} \int_{-u}^{u} (1 - \varphi(t)) dt < \varepsilon$.

(Note that the integral is real.) Since $\varphi_n(t) \to \varphi(t)$ and $|1 - \varphi_n(t)| \leq 2$ by Lebesgue’s dominated convergence theorem, there exists $n_0$ such that $\frac{1}{u} \int_{-u}^{u} (1 - \varphi_n(t)) dt < 2\varepsilon$ for all $n > n_0$.

Now we use Fubini’s theorem:

$$
\frac{1}{u} \int_{-u}^{u} (1 - \varphi_n(t)) dt = E \left( \int_{-u}^{u} \frac{1 - e^{itX_n}}{u} dt \right) = E \left( \int_{-u}^{u} \frac{1 - \cos tX_n}{u} dt \right) = E \left( 2 - \int_{-u}^{u} \cos tX_n dt \right)
$$

$$
= 2E \left( 1 - \frac{\sin(u|X_n|)}{u|X_n|} \right) \geq 2 \int_{|X_n| \geq 2} \left( 1 - \frac{\sin(u|X_n|)}{u|X_n|} \right) dP
$$

$$
\geq 2 \int_{|X_n| \geq 2} \left( 1 - \frac{1}{u|X_n|} \right) dP \geq P(|X_n| \geq 2).
$$

So with $K = 2/u$ we have $P(|X_n| \geq K) \leq 2\varepsilon$ for all $n > n_0$. Increasing $K$ if necessary we can ensure that $P(|X_n| \geq K) \leq 2\varepsilon$ holds also for the finitely many $n$ preceding $n_0$.

As an application we give another proof of Slutski’s theorem (Theorem 9.4).

**Corollary 10.15.** If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 0$ then $X_n + Y_n \xrightarrow{D} X$

**Proof.**

$$
|\varphi_{X_n+Y_n}(t) - \varphi_{X_n}(t)| = \left| E \left( e^{itX_n} (e^{itY_n} - 1) \right) \right| \leq E \left( |e^{itY_n} - 1| \right)
$$
Since $|e^{itY_n} - 1| \overset{P}{\to} 0$ by Exercise 4.25 and is bounded by 2, Lebesgue’s dominated convergence theorem (Theorem 6.9) shows that $\lim_{n \to \infty} \varphi_{X_n + Y_n}(t) = \varphi_{X_n}(t) = 0$. So both sequences have the same limit. \qed

---

**Required Exercises**

**Exercise 10.3.** Let $Z$ be the standard normal $N(0,1)$ r.v. Use Example 10.2 to compute the characteristic function of $X = \mu + \sigma Z$. (This is called the general normal r.v.)

**Exercise 10.4.** Let $X$ be a Poisson random variable with parameter $\lambda$. That is, $P(X = k) = e^{-\lambda} \lambda^k / k!$, $k = 0, 1, \ldots$. Compute the characteristic function of $(X - \lambda) / \sqrt{\lambda}$ and find its limit as $\lambda \to \infty$.

**Exercise 10.5.** Suppose $X_1, X_2, \ldots$ are i.i.d. with $P(X = \pm 1) = 1/2$. Let $S_n = X_1 + \cdots + X_n$. Compute the characteristic function of $\frac{1}{\sqrt{n}} S_n$ and find its limit as $n \to \infty$.

**Exercise 10.6.** Suppose $U_n$ are uniform on $(-n, n)$. Compute the characteristic function $\varphi_n(t)$ and find its limit as $n \to \infty$.

**Exercise 10.7.** Prove the case $n = 3$ of Theorem 10.7.

**Exercise 10.8.** Suppose $X, Y$ are independent exponential (i.e. with density $e^{-x}$ for $x > 0$). Compute the characteristic function of $X - Y$.

**Exercise 10.9.** Suppose $X_n \overset{D}{\to} X$ and $a_n \to a$, $b_n \to b$. Use characteristic functions to show that $a_n X_n + b_n \to aX + b$.

**Exercise 10.10.** Suppose $X_n \overset{D}{\to} X$, $Y_n \overset{D}{\to} Y$ and each pair $(X_n, Y_n)$ consists of independent random variables (on some probability space $\Omega_n$). Show that $X_n + Y_n \overset{D}{\to} S$ where the law of $S$ is the convolution of the laws of $X$ and $Y$.

---

**Additional Exercises**

**Exercise 10.11.** Suppose $X_1, X_2$ are independent and take values $\pm 1$ with equal probabilities. Show that the characteristic function of $X_1 + X_2$ is $\cos^2 t$.

**Exercise 10.12.** Show that

$$P(X = a) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita} \varphi(t) dt$$

**Exercise 10.13.** Suppose $P(X = x_k) > 0$. Show that

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |\varphi(t)|^2 dt = \sum_k (P(X = x_k))^2$$

*Hint* See [Billingsley, Exercise 26.13]
The Central Limit Theorem

1. Sums of independent identically distributed random variables

Denote by $Z$ the "standard normal random variable" with density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

**Lemma 11.1.** $e^{itZ} = e^{-t^2/2}$

**Proof.** We use the same calculation as for the moment generating function:

$$\int_{-\infty}^{\infty} \exp(itx - \frac{1}{2}x^2)dx = e^{-t^2/2} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x-\frac{it}{2})^2)dx = \sqrt{2\pi}$$

Note that $e^{-x^2/2}$ is an analytic function so $\oint e^{-z^2/2}dz = 0$ over any closed path. So

$$\int_{-A}^{A} \exp(-(x-it)^2/2)dx - \int_{-A}^{A} e^{-x^2/2}dx + \int_{0}^{it} \exp(-(A-is)^2/2)ds - \int_{0}^{it} \exp(-(A-is)^2/2)ds = 0$$

□

**Theorem 11.2** (CLT for i.i.d.). Suppose $\{X_n\}$ is i.i.d. with mean $m$ and variance $0 < \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n - nm}{\sigma\sqrt{n}} \xrightarrow{p} Z$$

This is one of the special cases of the Lindeberg theorem and the proof uses characteristic functions. Note that $\varphi_{S_n/\sqrt{n}}(t) = e^{-t^2/2}$ when $X_j$ are independent $N(0, 1)$.

In general, $\varphi_{S_n/\sqrt{n}}(t)$ is a complex number. For example, when $X_n$ are exponential with parameter $\lambda = 1$, the conclusion says that

$$\varphi_{S_n/\sqrt{n}}(t) = \frac{e^{-it\sqrt{n}}}{(1 - i\frac{t}{\sqrt{n}})^n} \rightarrow e^{-t^2/2}$$
which is not so obvious to see. On the other hand, characteristic function in Exercise 10.5 on page 119 is real and the limit can be found using calculus:

\[ \phi_{S_n/\sqrt{n}}(t) = \cos^n(t/\sqrt{n}) \to e^{-t^2/2}. \]

Here is a simple inequality that will suffice for the proof in the general case.

**Lemma 11.3.** If \( z_1, \ldots, z_m \) and \( w_1, \ldots, w_m \) are complex numbers of modulus at most 1 then

\[
|z_1 \ldots z_m - w_1 \ldots w_m| \leq \sum_{k=1}^m |z_k - w_k|
\]

**Proof.** Write the left hand side of (11.1) as a telescoping sum:

\[
z_1 \ldots z_m - w_1 \ldots w_m = \sum_{k=1}^m z_1 \ldots z_{k-1}(z_k - w_k)w_{k+1} \ldots w_m
\]

\(\square\)

Example 11.1. We show how to complete the proof for the exponential distribution.

\[
\left| \frac{e^{-it/\sqrt{n}}}{1 - i t / \sqrt{n}} - e^{-t^2/2} \right| = \left| \left( \frac{e^{-it/\sqrt{n}}}{1 - i t / \sqrt{n}} \right)^n - (e^{-t^2/(2n)})^n \right| 
\leq n \left| \frac{e^{-it/\sqrt{n}}}{1 - i t / \sqrt{n}} - e^{-t^2/(2n)} \right|
\]

\[
= n \left| \frac{1 - it/\sqrt{n} + t^2/(2n) + it^3/(6n\sqrt{n}) - \cdots - 1 + t^2/(2n) - t^4/(6n^2) + \cdots}{1 - i t / \sqrt{n}} \right|
\]

\[
= n \left| \left( 1 - \frac{it}{\sqrt{n}} - \frac{t^2}{2n} - \frac{it^3}{6n\sqrt{n}} + \cdots \right) \left( 1 + \frac{t}{\sqrt{n}} - \frac{t^2}{n} + \cdots \right) - \frac{t^2}{2n} - \frac{t^4}{6n^2} + \cdots \right|
\]

\[
= n \left| \left( 1 - \frac{t^2}{2n} + \frac{t^2}{2n} + \frac{t^3}{6n\sqrt{n}} - \cdots - 1 + \frac{t^2}{2n} - \frac{t^4}{6n^2} + \cdots \right) \leq n C(t) \right| \to 0.
\]

Proof of Theorem 11.2. Without loss of generality we may assume \( m = 0 \) and \( \sigma = 1 \). We have \( \phi_{S_n/\sqrt{n}}(t) = \phi_X(t/\sqrt{n})^n \). For a fixed \( t \in \mathbb{R} \) choose \( n \) large enough so that \( 1 - \frac{t^2}{2n} > -1 \). For such \( n \), we can apply (11.1) with \( z_k = \phi_X(t/\sqrt{n}) \) and \( w_k = 1 - \frac{t^2}{2n} \). We get

\[
\left| \phi_{S_n/\sqrt{n}}(t) - (1 - \frac{t^2}{2n})^n \right| \leq n \left| \phi_X(t/\sqrt{n}) - 1 - \frac{t^2}{2n} \right| \leq t^2 E \min \left\{ \frac{|t||X|^3}{\sqrt{n}}, X^2 \right\}
\]

Noting that \( \lim_{n \to \infty} \min \{ |t||X|^3/\sqrt{n}, X^2 \} = 0 \), by dominated convergence theorem (the integrand is dominated by the integrable function \( X^2 \)) we have \( E \min \left\{ \frac{|t||X|^3}{\sqrt{n}}, X^2 \right\} \to 0 \) as \( n \to \infty \). So

\[
\lim_{n \to \infty} \left| \phi_{S_n/\sqrt{n}}(t) - (1 - \frac{t^2}{2n})^n \right| = 0.
\]

It remains to notice that \( (1 - \frac{t^2}{2n})^n \to e^{-t^2/2} \). \( \square \)

**Remark 11.4.** If \( X_n \xrightarrow{D} Z \) then the cumulative distribution functions converge uniformly: \( \sup_n |P(X_n \leq x) - P(Z \leq x)| \to 0 \).
Example 11.2 (Normal approximation to Binomial). If $X_n$ is $Bin(n,p)$ and $p$ is fixed then $P\left(\frac{1}{\sqrt{n}}X_n < p + x/\sqrt{n}\right) \to P(Z \leq x/\sqrt{p(1-p)}$ as $n \to \infty$.

Example 11.3 (Normal approximation to Poisson). If $X_\lambda$ is $Poiss$ and $p$ is fixed then $(X_\lambda - \lambda)/\sqrt{\lambda} \xrightarrow{D} Z$ as $\lambda \to \infty$. (Strictly speaking, the CLT gives only convergence of $(X_{\lambda n} - \lambda n)/\sqrt{n\lambda} \xrightarrow{D} Z$ as $n \to \infty$.)

2. General form of a limit theorem

The general problem of convergence in distribution can be stated as follows: Given a sequence $Z_n$ of random variables, find normalizing constants $a_n, b_n$ and a limiting distribution/random variable $Z$ such that $(Z_n - b_n)/a_n \to Z$.

In Example 9.1, $Z_n$ is a maximum, $a_n = 1, b_n = \log n$.

In Theorem 11.2, $Z_n$ is the sum, the normalizing constants are $b_n = E(S_n)$ and $a_n = \sqrt{Var(S_n)}$, and we will make the same choice for sums of independent random variables in the next section. However, finding an appropriate normalization for CLT may be not obvious or easy, see Section 5.

One may wonder how much flexibility do we have in the choice of the normalizing constants $a_n, b_n$.

Theorem 11.5 (Convergence of types). Suppose $X_n \xrightarrow{D} X$ and $a_nX_n + b_n \xrightarrow{D} Y$ for some $a_n > 0, b_n \in \mathbb{R}$, and both $X, Y$ are non-degenerate. Then $a_n \to a > 0$ and $b_n \to b$ and in particular $Y$ has the same law as $aX + b$.

So if $(Z_n - b_n)/a_n \to Z$ and $(Z_n - b'_n)/a'_n \to Z'$ then $(Z_n - b'_n)/a'_n = \frac{a_n}{a'_n}((Z_n - b_n)/a_n) + (b_n - b'_n)/a'_n$, which means that $a_n/a'_n \to a > 0$ and $(b_n - b'_n)/a'_n \to b$. So $a'_n = a_n/a, b'_n = b_n - \frac{b}{a}a_n$ and $Z' = aZ + b$.

Proof. To be written... □

It is clear that independence alone is not sufficient for the CLT.

3. Lindeberg’s theorem

The setting is of sums of triangular arrays: For each $n$ we have a family of independent random variables

$$X_{n1} \ldots X_{nr_n}$$

and we set $S_n = X_{n1} + \ldots + X_{nr_n}$.

For Theorem 11.2, the triangular array can be $X_{nk} = \frac{X_k - m}{\sigma \sqrt{r}}$. Or one can take $X_{nk} = \frac{X_k - m}{\sigma}$...

Through this section we assume that random variables are square-integrable with mean zero, and we use the notation

$$E(X_{nk}) = 0, \sigma_{nk}^2 = E(X_{nk}^2), s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$$

(11.2)
**Definition 11.1** (The Lindeberg condition). We say that the Lindeberg condition holds if

\[
\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP = 0 \text{ for all } \varepsilon > 0
\]

(11.3)

(Note that strict inequality \(\int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP\) can be replaced by \(\int_{|X_{nk}| \geq \varepsilon s_n} X_{nk}^2 dP\) and the resulting condition is the same.)

**Remark 11.6.** Under the Lindeberg condition, we have

\[
\lim_{n \to \infty} \max_{1 \leq k \leq r_n} \frac{\sigma_{nk}^2}{s_n^2} = 0
\]

Indeed,

\[
\sigma_{nk}^2 = \int_{|X_{nk}| \leq \varepsilon s_n} X_{nk}^2 dP + \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP \leq \varepsilon s_n^2 + \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP
\]

So

\[
\max_{1 \leq k \leq r_n} \frac{\sigma_{nk}^2}{s_n^2} \leq \varepsilon + \frac{1}{s_n^2} \max_{1 \leq k \leq r_n} \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP \leq \varepsilon + \frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP
\]

**Theorem 11.7** (Lindeberg CLT). Suppose that for each \(n\) the sequence \(X_{n1} \ldots X_{nr_n}\) is independent with mean zero. If the Lindeberg condition holds for all \(\varepsilon > 0\) then \(S_n/s_n \xrightarrow{D} Z\).

**Example 11.4** (Proof of Theorem 11.2). In the setting of Theorem 11.2, we have \(X_{nk} = \frac{X_k - m}{\sigma}\) and \(s_n = \sqrt{n}\). The Lindeberg condition is

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{r_n} \int_{|X_k - m| > \varepsilon \sigma \sqrt{n}} \frac{(X_k - m)^2}{\sigma^2} dP = \lim_{n \to \infty} \frac{1}{\sigma^2} \int_{|X_1 - m| > \varepsilon \sigma \sqrt{n}} (X_1 - m)^2 dP = 0
\]

by Lebesgue dominated convergence theorem, say. (Or by Corollary 6.12 on page 73.)

**Proof.** Without loss of generality we may assume that \(s_n^2 = 1\) so that \(\sum_{k=1}^{r_n} \sigma_{nk}^2 = 1\). Denote \(\varphi_{nk} = E(e^{itX_{nk}})\). From (10.13) we have

\[
(11.5) \quad \left| \varphi_{nk}(t) - (1 - \frac{1}{2} t^2 \sigma_{nk}^2) \right| \leq E \left( \min\{|tX_{nk}|^2, |tX_{nk}|^3\} \right)
\]

\[
\leq \int_{|X_{nk}| < \varepsilon} |tX_{nk}|^3 dP + \int_{|X_{nk}| \geq \varepsilon} |tX_{nk}|^2 dP \leq t^3 \varepsilon \sigma_{nk}^2 + t^2 \int_{|X_{nk}| \geq \varepsilon} X_{nk}^2 dP
\]

Using (11.1), we see that

\[
(11.6) \quad \left| \varphi_{Sn}(t) - \prod_{k=1}^{n} \left(1 - \frac{1}{2} t^2 \sigma_{nk}^2\right) \right| \leq \varepsilon t^3 \sum_{k=1}^{n} \sigma_{nk}^2 + t^2 \sum_{k=1}^{n} \int_{|X_{nk}| \geq \varepsilon} |X_{nk}^2| dP
\]

This shows that

\[
\lim_{n \to \infty} \left| \varphi_{Sn}(t) - \prod_{k=1}^{n} \left(1 - \frac{1}{2} t^2 \sigma_{nk}^2\right) \right| = 0
\]

It remains to verify that \(\lim_{n \to \infty} e^{-t^2/2} - \prod_{k=1}^{n} \left(1 - \frac{1}{2} t^2 \sigma_{nk}^2\right) = 0\).
To do so, we apply the previous proof to the triangular array $\sigma_{n,k}Z_k$ of independent normal random variables. Note that

$$\varphi \sum Z_{nk}(t) = \prod_{k=1}^{r_n} e^{-t^2 \sigma_{nk}^2 / 2} = e^{-t^2 / 2}$$

We only need to verify the Lindeberg condition for $\{Z_{nk}\}$:

$$\int_{|Z_{nk}| > \varepsilon} Z_{nk}^2 dP = \sum_{k=1}^{r_n} \sigma_{nk}^2 \int_{|x| > \varepsilon / \sigma_{nk}} x^2 f(x) dx$$

So

$$\sum_{k=1}^{r_n} \int_{|Z_{nk}| > \varepsilon} Z_{nk}^2 dP \leq \sum_{k=1}^{r_n} \sigma_{nk}^2 \max_{1 \leq k \leq r_n} \int_{|x| > \varepsilon / \sigma_{nk}} x^2 f(x) dx$$

The right hand side goes to zero as $n \to \infty$, because by $\max_{1 \leq k \leq r_n} \sigma_{nk} \to 0$ by (11.4).

4. Lyapunov’s theorem

**Theorem 11.8.** Suppose that for each $n$ the sequence $X_{n1} \ldots X_{nr_n}$ is independent with mean zero. If the Lyapunov’s condition

$$\lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E|X_{nk}|^{2+\delta} = 0$$

holds for some $\delta > 0$, then $S_n / s_n \xrightarrow{D} Z$.

**Proof.** We use the following bound to verify Lindeberg’s condition:

$$\frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}| > \varepsilon s_n} X_{nk}^2 dP \leq \frac{1}{\varepsilon^2 s_n^{2+\delta}} \sum_{k=1}^{r_n} \int_{|X_{nk}| > \varepsilon s_n} |X_{nk}|^{2+\delta} dP \leq \frac{1}{\varepsilon^2 s_n^{2+\delta}} \sum_{k=1}^{n} E|X_{nk}|^{2+\delta} \xrightarrow{\text{by (11.4)}} 0$$

**Corollary 11.9.** Suppose $X_k$ are independent with mean zero, variance $\sigma^2$ and that $\sup_k E|X_k|^{2+\delta} < \infty$. Then $S_n / \sqrt{n} \xrightarrow{D} \sigma Z$.

**Proof.** Let $C = \sup_k E|X_k|^{2+\delta}$. Then $s_n = \sqrt{n}$ and $\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E(|X_k|^{2+\delta}) \leq C / n^{\delta/2} \to 0$, so Lyapunov’s condition is satisfied.

**Corollary 11.10.** Suppose $X_k$ are independent, uniformly bounded, and have mean zero. If $\sum_n \text{Var}(X_n) = \infty$, then $S_n / \sqrt{\text{Var}(S_n)} \xrightarrow{D} N(0,1)$.

**Proof.** Suppose $|X_n| \leq C$ for a constant $C$. Then

$$\frac{1}{s_n^2} \sum_{k=1}^{n} E|X_n|^3 \leq \frac{C}{s_n^2} \xrightarrow{s_n \to \infty} 0$$
5. Normal approximation without Lindeberg condition

One basic idea is truncation: \( X_n = X_n I_{|X_n| \leq a_n} + X_n I_{|X_n| > a_n} \). One wants to show that \( \frac{1}{s_n} \sum X_k I_{|X_k| \leq a_n} \to Z \) and that \( \frac{1}{s_n} \sum X_k I_{|X_k| > a_n} \to 0 \). Then \( S_n/s_n \) is asymptotically normal by Slutski’s theorem.

**Example 11.5.** Let \( X_1, X_2, \ldots \) be independent random variables with the distribution \((k \geq 1)\)

\[
\begin{align*}
\Pr(X_k &= \pm 1) = 1/4, \\
\Pr(X_k &= k^k) = 1/4^k, \\
\Pr(X_k &= 0) = 1/2 - 1/4^k.
\end{align*}
\]

Then \( \sigma_k^2 = \frac{1}{2} + \left(\frac{k}{4}\right)^k \) and \( s_n \geq n^{n/4^k} \). But \( S_n/s_n \to 0 \) and in fact we have \( S_n/\sqrt{n} \to Z/\sqrt{2} \). To see this, note that \( Y_k = X_k I_{|X_k| \leq 1} \) are independent with mean 0, variance \( \frac{1}{2} \) and \( P(Y_k \neq X_k) = 1/4^k \) so by the first Borel Cantelli Lemma (Theorem 3.8) \( \frac{1}{\sqrt{n}} \sum_{k=1}^n (Y_k - X_k) \leq \frac{U}{\sqrt{n}} \to 0 \) with probability one.

It is sometimes convenient to use Corollary 9.5 (Exercise 9.2) combined with the law of large numbers. This is how one needs to proceed in Exercise 11.2.

**Example 11.6.** Suppose \( X_1, X_2, \ldots \), are i.i.d. with mean 0 and variance \( \sigma^2 > 0 \). Then

\[
\frac{\sum_{k=1}^n X_k}{\sqrt{\sum_{k=1}^n X_k^2}}
\]

converges in distribution to \( N(0, 1) \). To see this, write

\[
\frac{\sum_{k=1}^n X_k}{\sqrt{\sum_{k=1}^n X_k^2}} = \frac{\sigma}{\sqrt{1/\sum_{k=1}^n X_k^2}} \times \frac{\sum_{k=1}^n X_k}{\sigma \sqrt{n}}
\]

and note that the first factor converges to 1 with probability one.

---

**Required Exercises**

**Exercise 11.1.** Suppose \( a_{nk} \) is an array of numbers such that \( \sum_{k=1}^n a_{nk}^2 = 1 \) and \( \max_{1 \leq k \leq n} |a_{nk}| \to 0 \). Let \( X_j \) be i.i.d. with mean zero and variance 1. Show that \( \sum_{k=1}^n a_{nk} X_k \to Z \).

**Exercise 11.2.** Suppose that \( X_1, X_2, \ldots \) are i.i.d., \( E(X_1) = 1 \), \( E(X_1^2) = \sigma^2 < \infty \). Let \( \bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j \). Show that for all \( k > 0 \)

\[
\sqrt{n} \left( \bar{X}_n^k - 1 \right) \to N(0, k\sigma)
\]

as \( n \to \infty \).

**Exercise 11.3.** Suppose \( X_1, X_2, \ldots \) are independent, \( X_k = \pm 1 \) with probability \( \frac{1}{2}(1 - k^{-2}) \) and \( X_k = \pm k \) with probability \( \frac{1}{2}k^{-2} \). Let \( S_n = \sum_{k=1}^n X_k \)

(i) Show that \( S_n/\sqrt{n} \to N(0, 1) \)
(ii) Is the Lindeberg condition satisfied?

**Exercise 11.4.** Suppose $X_1, X_2, \ldots$ are independent random variables with distribution $\Pr(X_k = 1) = p_k$ and $\Pr(X_k = 0) = 1 - p_k$. Prove that if $\sum \text{Var}(X_k) = \infty$ then

$$\frac{\sum_{k=1}^{n}(X_k - p_k)}{\sqrt{\sum_{k=1}^{n}p_k(1 - p_k)}} \xrightarrow{D} N(0, 1).$$

**Exercise 11.5.** Suppose $X_k$ are independent and have density $\frac{1}{|x|^3}$ for $|x| > 1$. Show that $S_n \sqrt{n \log n} \xrightarrow{D} N(0, 1)$.

*Hint:* Verify that Lyapunov’s condition (11.7) holds with $\delta = 1$ for truncated random variables. Several different truncations can be used, but technical details differ:

- $Y_k = X_k I_{|X_k| \leq \sqrt{k}}$ is a solution in [Billingsley]. To show that $\frac{1}{\sqrt{n \log n}} \sum_{k=1}^{n}(X_k - Y_k) \xrightarrow{P} 0$ use $L_1$-convergence.
- Triangular array $Y_{nk} = X_k I_{|X_k| \leq \sqrt{n}}$ is simpler computationally
- Truncation $Y_k = X_k I_{|X_k| \leq \sqrt{k \log k}}$ leads to “asymptotically equivalent” sequences.

---

**Some previous prelim problems**

**Exercise 11.6** (May 2018). (To be added latter)

**Exercise 11.7** (Aug 2017). Let $\{X_n\}_{n \in \mathbb{N}}$ be a collection of independent random variables with

$$\Pr(X_n = \pm n^2) = \frac{1}{2n^{\beta}} \quad \text{and} \quad \Pr(X_n = 0) = 1 - \frac{1}{n^{\beta}}, \quad n \in \mathbb{N},$$

where $\beta \in (0, 1)$ is fixed for all $n \in \mathbb{N}$. Consider $S_n := X_1 + \cdots + X_n$. Show that

$$\frac{S_n}{n^{\gamma}} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

for some $\sigma > 0$, $\gamma > 0$. Identify $\sigma$ and $\gamma$ as functions of $\beta$. You may use the formula

$$\sum_{k=1}^{n}k^{\theta} \sim \frac{n^{\theta+1}}{\theta + 1}$$

for $\theta > 0$, and recall that by $a_n \sim b_n$ we mean $\lim_{n \to \infty} a_n/b_n = 1$.

**Exercise 11.8** (May 2017). Let $\{X_n\}_{n \in \mathbb{N}}$ be independent random variables with $\Pr(X_n = 1) = 1/n = 1 - \Pr(X_n = 0)$. Let $S_n := X_1 + \cdots + X_n$ be the partial sum.

(i) Show that

$$\lim_{n \to \infty} \frac{\mathbb{E}S_n}{\log n} = 1 \quad \text{and} \quad \lim_{n \to \infty} \frac{\text{Var}(S_n)}{\log n} = 1.$$

(ii) Prove that

$$\frac{S_n - \log n}{\sqrt{\log n}} \xrightarrow{D} \mathcal{N}(0, 1).$$
as \( n \to \infty \). Explain which central limit theorem you use. State and verify all the conditions clearly.

Hint: recall the relation \( \lim_{n \to \infty} \sum_{k=1}^{n} 1/k \log n = 1 \).

**Exercise 11.9** (May 2016). (a) State Lindeberg–Feller central limit theorem.

(b) Use Lindeberg–Feller central limit theorem to prove the following. Consider a triangular array of random variables \( \{Y_{n,k}\}_{n \in \mathbb{N}, k=1, \ldots, n} \) such that for each \( n \), \( \mathbb{E}Y_{n,k} = 0 \), \( k = 1, \ldots, n \), and \( \{Y_{n,k}\}_{k=1, \ldots, n} \) are independent. In addition, with \( \sigma_n := \left( \sum_{k=1}^{n} \mathbb{E}Y_{n,k}^2 \right)^{1/2} \), assume that
\[
\lim_{n \to \infty} \frac{1}{\sigma_n^4} \sum_{k=1}^{n} \mathbb{E}Y_{n,k}^4 = 0.
\]

Show that \( Y_{n,1} + \cdots + Y_{n,n} \sigma_n \xrightarrow{D} \mathcal{N}(0,1) \).

**Exercise 11.10** (Aug 2015). Let \( \{U_n\}_{n \in \mathbb{N}} \) be a collection of i.i.d. random variables with \( \mathbb{E}U_n = 0 \) and \( \mathbb{E}U_n^2 = \sigma^2 \in (0, \infty) \). Consider random variables \( \{X_n\}_{n \in \mathbb{N}} \) defined by \( X_n = U_n + U_{2n}, n \in \mathbb{N} \), and the partial sum \( S_n = X_1 + \cdots + X_n \). Find appropriate constants \( \{a_n, b_n\}_{n \in \mathbb{N}} \) such that
\[
\frac{S_n - b_n}{a_n} \xrightarrow{D} \mathcal{N}(0,1).
\]

**Exercise 11.11** (May 2015). Let \( \{U_n\}_{n \in \mathbb{N}} \) be a collection of i.i.d. random variables distributed uniformly on interval \((0,1)\). Consider a triangular array of random variables \( \{X_{n,k}\}_{k=1, \ldots, n, n \in \mathbb{N}} \) defined as
\[
X_{n,k} = 1_{\{\sqrt{n}U_k \leq 1\}} - \frac{1}{\sqrt{n}}.
\]
Find constants \( \{a_n, b_n\}_{n \in \mathbb{N}} \) such that
\[
\frac{X_{n,1} + \cdots + X_{n,n} - b_n}{a_n} \xrightarrow{D} \mathcal{N}(0,1).
\]

**Exercise 11.12** (Aug 2014). Let \( X_1, X_2, \ldots \) be independent and identically distributed random variables with
\[
P(X_i = 1) = P(X_i = -1) = 1/2.
\]
Prove that
\[
\sqrt{3} \sum_{k=1}^{n} kX_k \xrightarrow{D} \mathcal{N}(0,1)
\]
(You may use formulas \( \sum_{j=1}^{n} j^2 = \frac{1}{6}n(n+1)(2n+1) \) and \( \sum_{j=1}^{n} j^3 = \frac{1}{4}n^2(n+1)^2 \) without proof.)

**Exercise 11.13** (May 2014). Let \( \{X_{nk} : k = 1, \ldots, n, n \in \mathbb{N}\} \) be a family of independent random variables satisfying
\[
P \left( X_{nk} = \frac{k}{\sqrt{n}} \right) = P \left( X_{nk} = -\frac{k}{\sqrt{n}} \right) = P(X_{nk} = 0) = 1/3
\]
Let \( S_n = X_{n1} + \cdots + X_{nn} \). Prove that \( S_n/s_n \) converges in distribution to a standard normal random variable for a suitable sequence of real numbers \( s_n \).
Some useful identities:
\[
\begin{align*}
\sum_{k=1}^{n} k &= \frac{1}{2} n(n + 1) \\
\sum_{k=1}^{n} k^2 &= \frac{1}{6} n(n + 1)(2n + 1) \\
\sum_{k=1}^{n} k^3 &= \frac{1}{4} n^2 (n + 1)^2 
\end{align*}
\]

**Exercise 11.14** (Aug 2013). Suppose \(X_1, Y_1, X_2, Y_2, \ldots\) are independent identically distributed with mean zero and variance 1. For integer \(n\), let
\[
U_n = \frac{1}{n} \left( \sum_{j=1}^{n} X_j \right)^2 + \frac{1}{n} \left( \sum_{j=1}^{n} Y_j \right)^2.
\]
Prove that \(\lim_{n \to \infty} P(U_n \leq u) = 1 - e^{-u/2}\) for \(u > 0\).

**Exercise 11.15** (May 2013). Suppose \(X_{n,1}, X_{n,2}, \ldots\) are independent random variables centered at expectations (mean 0) and set \(s_n^2 = \sum_{k=1}^{n} \mathbb{E} \left( (X_{n,k})^2 \right)\). Assume for all \(k\) that \(|X_{n,k}| \leq M_n\) with probability 1 and that \(M_n/s_n \to 0\). Let \(Y_{n,i} = 3X_{n,i} + X_{n,i+1}\). Show that \(\frac{Y_{n,1} + Y_{n,2} + \ldots + Y_{n,n}}{s_n}\) converges in distribution and find the limiting distribution.
Chapter 12

Limit Theorems in $\mathbb{R}^k$

This is based on [Billingsley, Section 29]

1. The basic theorems

If $X : \Omega \rightarrow \mathbb{R}^k$ is measurable, then $X$ is called a random vector. $X$ is also called a $k$-variate random variable, as $X = (X_1, \ldots, X_k)$.

Recall that a probability distribution of $X$ is a probability measure $\mu$ on Borel subsets of $\mathbb{R}^k$ defined by

$\mu(U) = P(\{\omega : X(\omega) \in U\})$.

Recall that a (joint) cumulative distribution function of $X = (X_1, \ldots, X_n)$ is a function $F : \mathbb{R}^n \rightarrow [0, 1]$ such that

$F(x_1, \ldots, x_k) = P(X_1 \leq x_1, \ldots, X_k \leq x_k)$

From $\pi - \lambda$ theorem we know that $F$ determines uniquely $\mu$. In particular, if

$F(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_k} f(y_1, \ldots, y_k)dy_1 \ldots dy_k$

then $\mu(U) = \int_U f(y_1, \ldots, y_k)dy_1 \ldots dy_k$.

Let $X : \Omega \rightarrow \mathbb{R}^k$ be a sequence of random vectors.

Definition 12.1. We say that $X_n$ converges in distribution to $X$ if for every bounded continuous function $f : \mathbb{R}^k \rightarrow \mathbb{R}$ the sequence of numbers $E(f(X_n))$ converges to $Ef(X)$.

We will write $X_n \xrightarrow{D} X$; if $\mu_n$ is the law of $X_n$ we will also write $\mu_n \rightarrow D$; the same notation in the language of cumulative distribution functions is $F_n \xrightarrow{D} F$; the latter can be defined as $F_n(x) \xrightarrow{D} F(x)$ for all points of continuity of $F$, but it is simpler to use Definition 12.1.

Proposition 12.1. If $X_n \xrightarrow{D} X$ and $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is a continuous function then $g(X_n) \xrightarrow{D} g(X)$

For example, if $(X_n, Y_n) \xrightarrow{D} (Z_1, Z_2)$ then $X_n^2 + Y_n^2 \xrightarrow{D} Z_1^2 + Z_2^2$.

Proof. Denoting by $Y_n = g(X_n)$, we see that for any bounded continuous function $f : \mathbb{R}^m \rightarrow \mathbb{R}$, $f(bY_n)$ is a bounded continuous function $f \circ g$ of $X_n$. \qed
Theorem 12.2. For a sequence $\mu_n$ of probability measures on the Borel sets of $\mathbb{R}^k$, the following are equivalent:

(i) $\mu_n \overset{D}{\to} \mu$
(ii) $\limsup_{n \to \infty} \mu_n(C) \leq \mu(C)$ for all closed sets $C \subset \mathbb{R}^k$.
(iii) $\liminf_{n \to \infty} \mu_n(G) \leq \mu(G)$ for all open sets $G \subset \mathbb{R}^k$.
(iv) $\lim_{n \to \infty} \mu_n(A) = \mu(A)$ for all sets $A \subset \mathbb{R}^k$ such that $\mu(\partial A) = 0$

Proof. The detailed proof is omitted. Here are some steps:

- By passing to complements, it is clear that (2) and (3) are equivalent.
- Since the interior $A^o$ of a set $A$ is its subset, $A^o \subset A \subset \bar{A}$. So $\mu_n(A^o) \leq \mu_n(A) \leq \mu_n(\bar{A})$ and we get $\mu(A^o) \leq \liminf_n \mu_n(A) \leq \limsup_n \mu_n(A) \leq \mu(\bar{A})$
  Since $\partial A = \bar{A} \setminus A^o$, we have $\mu(A^o) = \mu(\bar{A}) = \mu(A)$ so it is clear that (2)+(3) imply (4).
- To see how (1) implies (2), fix closed set $C \subset \mathbb{R}^k$ and consider a bounded continuous function $f$ such that $f = 1$ on $C$ and $f = 0$ on $C^c = \{x \in \mathbb{R}^k : d(x,C) \leq \epsilon\}$. Then $\mu_n(C) \leq \int f(x) \mu_n(dx) \to \int f \mu(dx) \leq \mu(C^c)$. Since $\lim_{\epsilon \to 0} \mu(C^c) = \mu(\bigcap_{\epsilon > 0} C_{\epsilon}) = \mu(C)$, we get the conclusion.

\[\square\]

Definition 12.2. The sequence of measures $\mu_n$ on $\mathbb{R}^k$ is tight if for every $\epsilon > 0$ there exists a compact set $K \subset \mathbb{R}^k$ such that $\mu_n(K) \geq 1 - \epsilon$ for all $n$.

Theorem 12.3. If $\mu_n$ is a tight sequence of probability measures then there exists $\mu$ and a subsequence $\mu_{n_k}$ such that $\mu_{n_k} \overset{D}{\to} \mu$.

Proof. The detailed proof is omitted.

Here are the main steps in the proof:

Corollary 12.4. If $\{\mu_n\}$ is a tight sequence of probability measures on Borel subsets of $\mathbb{R}^k$ and if each convergent subsequence has the same limit $\mu$, then $\mu_n \overset{D}{\to} \mu$.

2. Multivariate characteristic function

Recall the dot-product $\mathbf{x} \cdot \mathbf{y} := \mathbf{x}' \mathbf{y} \sum_{j=1}^{k} x_j y_j$. The multivariate characteristic function $\varphi : \mathbb{R}^k \to \mathbb{C}$ is

\[\varphi(t) = \mathbb{E} \exp(it \cdot \mathbf{X})\]

This is also written as $\varphi(t_1, \ldots, t_k) = \mathbb{E} \exp(\sum_{j=1}^{k} t_j X_j)$.

The inversion formula shows how to determine $\mu(U)$ for a rectangle $U = (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_k, b_k]$ such that $\mu(\partial U) = 0$:

\[\mu(U) = \lim_{T \to \infty} \frac{1}{(2\pi)^k} \int_{-T}^{T} \cdots \int_{-T}^{T} \prod_{j=1}^{k} e^{-i a_j t_j} - e^{-i b_j t_j} \frac{\varphi(t_1, \ldots, t_k)}{it_j} dt_1 \cdots dt_k\]

Thus the characteristic function determines the probability measure $\mu$ uniquely.

Corollary 12.5 (Cramer-Wold devise). The law of $\mathbf{X}$ is uniquely determined by the univariate laws $t \cdot \mathbf{X} = \sum_{j=1}^{k} t_j X_j$. 
Corollary 12.6. $X, Y$ are independent iff $\varphi_{X,Y}(s,t) = \varphi_X(s)\varphi_Y(t) $

Theorem 12.7. $X_n \overset{D}{\to} Y$ iff $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}^k$.

Example 12.1. If $X, Y$ are independent normal then $X + Y$ and $X - Y$ are independent normal. Indeed, $\varphi_{X+Y,X-Y}(s,t) = \varphi_X(s+t)\varphi_Y(s-t) = \exp((t+s)^2/2 + (s-t)^2/2) = e^{s^2}e^{t^2}$, and $\varphi_{X-Y}(s) = e^{s^2}e^{s^2} = e^{2s^2}$.

Corollary 12.8. If $Z_1, \ldots, Z_m$ are independent normal and $X = AZ$ then $\sum_j t_j X_j$ is (univariate) normal.

Proof. Let's simplify the calculations by assuming $Z_j$ are standard normal. The characteristic function of $S = \sum_j t_j X_j$ is

$$\varphi(s) = E \exp(ist \cdot X) = E \exp(ist \cdot AZ) = E \exp(is(A^T t) \cdot Z) = \prod_{i=1}^m e^{-s^2(A^T t)_i^2/2} = e^{-s^2\|A^T t\|^2/2} \Box$$

The generalization of this property is the simplest definition of the multivariate normal distribution. Note that $\|A^T t\|^2 = (A^T t) \cdot (A^T t) = t'AA^T t = t'\Sigma t$.

3. Multivariate normal distribution

Two equivalent approaches

Definition 12.3. $X$ is multivariate normal if there is a vector $m$ and a positive-definite matrix $\Sigma$ such that its characteristic function is

$$\varphi(t) = \exp \left( im^t t - \frac{1}{2} t'\Sigma t \right)$$

Notation: $N(m, \Sigma)$. (How do we know that this is a characteristic function? See the proof of Corollary 12.8!)

We need to show that this is indeed a characteristic function! But if it is, then by differentiation the parameters have interpretation: $E X = m$ and $\Sigma_{i,j} = \text{cov}(X_i, X_j)$.

Remark 12.9. If $X$ is normal $N(m, \Sigma)$, then $X - m$ is centered normal $N(0, \Sigma)$. In the sequel, to simplify notation we only discuss centered case.

The simplest way to define the univariate distribution is to start with a standard normal random variable $Z$ with density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, and then define the general normal as the linear function $X = \mu + \sigma Z$. It is then easy to work out the density of $X$ and the characteristic function, which is $\varphi_X(t) = e^{it\mu + \frac{1}{2}t^2\sigma^2}$.

Exercise: suppose $X, Y$ are independent normal. Use Definition 12.3 to verify that each of the following is bivariate normal: $X = (X, X), X = (X, Y), X = (X + \varepsilon Y, X - \varepsilon Y)$.
In $\mathbb{R}^k$ the role of the standard normal distribution is played by the distribution of $Z = (Z_1, \ldots, Z_k)$ of i.i.d. $N(0, 1)$ r.v.'s. Their density is

$$f(x) = \frac{1}{(2\pi)^{k/2}} e^{-\|x\|^2/2}$$  \hfill (12.3)

The characteristic function $Ee^{it'Z}$ is just the product of the individual characteristic functions $\prod_{j=1}^n e^{-t_j^2/2}$ which in vector notation is

$$\varphi_Z(y) = e^{-\|t\|^2/2}$$

**Definition 12.4.** We will say that $X$, written as a column vector, has multivariate normal distribution if $X = m + AZ$.

Clearly, $E(X) = m$. In the sequel we will only consider centered multivariate normal distribution with $E(X) = 0$.

**Remark 12.10.** Denoting by $a_k$ the columns of $A$, we have $X = \sum_{j=1}^k Z_j a_j$. This is the general feature of Gaussian vectors - they can be written as linear combinations of deterministic vectors with independent real-valued "noises" as coefficients.

**Proposition 12.11.** The characteristic function of the centered normal distribution is

$$\varphi(t) = \exp \left( -\frac{1}{2} t' \Sigma t \right)$$  \hfill (12.4)

where $\Sigma$ is a $k \times k$ positive definite matrix.

**Proof.** This is just a calculation:

$$Ee^{it'X} = Ee^{it'AZ} = Ee^{i(A't)'Z} = e^{-\|A't\|^2/2} = \exp \left( -\frac{1}{2} (A't)'A't \right) = \exp \left( -\frac{1}{2} t' A A' t \right) = \exp \left( -\frac{1}{2} t' \Sigma t \right)$$

\[\square\]

**Remark 12.12.** Notice that $E(XX') = E(AZZ'A') = AE(ZZ')A' = AIA' = \Sigma$ is the covariance matrix of $X$.

**Remark 12.13.** From linear algebra, any positive definite matrix $\Sigma = U\Lambda U'$ so each such matrix can be written as $\Sigma = AA'$ with $A = U\Lambda^{1/2}U'$. So $\varphi(t) = \exp(-\frac{1}{2} t' \Sigma t)$ is a characteristic function of $X = AZ$.

**Remark 12.14.** If $\det(\Sigma) > 0$ then $\det A \neq 0$ and (by linear algebra) the inverse $A^{-1}$ exists. The density of $X$ is recalculated from (12.3) as follows

$$f(x) = \frac{1}{(2\pi)^{k/2} \det(A)} e^{-\|A^{-1}x\|^2/2} = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2} x' \Sigma x}$$

Matrix $A$ in the representation $X = AZ$ is not unique, but the covariance matrix $\Sigma = AA'$ is unique.

**Example 12.2.** Suppose $\varphi(s, t) = e^{-s^2/2-t^2/2-\rho st}$. Then

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$
is non-negative definite for any $|\rho| \leq 1$ and this is a characteristic function of a random variable $X = (X_1, X_2)$ with univariate $N(0,1)$ laws, with correlation $E(X_1 X_2) = -\frac{\partial^2}{\partial s \partial t} \varphi(s, t) |_{s=t=0} = \rho$. If $Z_1, Z_2$ are independent $N(0,1)$ then

$$(12.5) \quad X_1 = Z_1, \quad X_2 = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$$

will have exactly the same second moments, and the same characteristic function.

Since $\det \Sigma = 1 - \rho^2$, when $\rho^2 \neq 1$ the matrix is invertible and the resulting bivariate normal density is

$$f(x, y) = \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( -\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)} \right)$$

From (12.5) we also see that $X_2 - \rho X_1$ is independent of $X_1$ and has variance $1 - \rho^2$

4. The CLT

**Theorem 12.15.** Let $X_n = (X_{n1}, \ldots, X_{nk})$ be independent random vectors with the same distribution and finite second moments. Denote $m = E X_k$ and $S_n = X_1 + \cdots + X_n$. Then

$$(S_n - nm)/\sqrt{n} \overset{D}{\to} Y$$

where $Y$ is a centered normal distribution with the covariance matrix $\Sigma = E(X_n X_n') - nm'$. 

**Proof.** Without loss of generality we can assume $m = 0$. Let $t \in \mathbb{R}^k$. Then $X_n := t'X_n$ are independent random variables with mean zero and variance $\sigma^2 = t' \Sigma t$. By Theorem 11.2, we have $S_n/\sqrt{n} \overset{D}{\to} \sigma Z$.

If $Y = (Y_1, \ldots, Y_k)$ has multivariate normal distribution with covariance $\Sigma$, then $t'Y$ is univariate normal with variance $\sigma^2$. So we showed that $t'S_n/\sqrt{n} \overset{D}{\to} t'Y$ for all $t \in \mathbb{R}^k$. This ends the proof by Theorem 12.7.

**Example 12.3.** Suppose $\xi_k, \eta_k$ are i.i.d with mean zero variance one. Then $(\sum_{k=1}^n \eta_k, \sum_{k=1}^n (\xi_k + \eta_k)) \overset{D}{\to} (Z_1, Z_1 + Z_2)$.

Indeed, random vectors $X_k = \begin{bmatrix} \xi_k \\ \xi_k + \eta_k \end{bmatrix}$ has covariance matrix $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ and $\begin{bmatrix} Z_1 \\ Z_1 + Z_2 \end{bmatrix}$ has the same covariance matrix.

4.1. Application: Chi-Squared test for multinomial distribution. A multinomial experiment has $k$ outcomes with probabilities $p_1, \ldots, p_k$. A multinomial random variable $S_n = (S_1(n), \ldots, S_k(n))$ lists observed counts per category in $n$ repeats of the multinomial experiment.

The following result is behind the use of the chi-squared statistics in tests of consistency.

**Theorem 12.16.** $\sum_{j=1}^k \frac{(S_{j}(n) - np_j)^2}{np_j} \overset{D}{\to} Z_1^2 + \cdots + Z_{k-1}^2$

**Proof.** Let’s prove this for $k = 3$. Consider independent random vectors $X_k$ that take three values $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ with probabilities $p_1, p_2, p_3$. Then $S_n$ is the sum of $n$ independent identically distributed vectors $X_1, \ldots, X_n$. 
Clearly, $E \mathbf{X}_k = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$. To compute the covariance matrix, write $\mathbf{X}$ for $\mathbf{X}_k$. For non-centered vectors, the covariance is $E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X}')$. We have

$$E(\mathbf{X}\mathbf{X}') = p_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \times [1 \ 0 \ 0] + p_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times [0 \ 1 \ 0] + p_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \times [0 \ 0 \ 1] = \begin{bmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{bmatrix}$$

So

$$\Sigma = E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X}') = \begin{bmatrix} p_1(1 - p_1) & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & p_2(1 - p_2) & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & p_3(1 - p_3) \end{bmatrix}$$

Then $\mathbf{S}_n$ is the sum of $n$ independent vectors, and the central limit theorem implies that

$$\frac{1}{\sqrt{n}} \left( \mathbf{S}_n - n \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \right) \xrightarrow{D} \mathbf{X}.$$

In particular, by Proposition 12.1 we have

$$\sum_{j=1}^{3} \frac{(S_j(n) - np_j)^2}{np_j} \xrightarrow{D} \sum_{j=1}^{3} \frac{X_j^2}{p_j}$$

where $(X_1, X_2, X_3)$ is multivariate normal with covariance matrix $\Sigma$.

Note that since $\sum_{j=1}^{k} S_j(n) = n$, the gaussian distribution is degenerate: $X_1 + X_2 + X_3 = 0$.

It remains to show that $\sum_{j=1}^{3} \frac{X_j^2}{p_j}$ has the same law as $Z_1^2 + Z_2^2$ i.e. that is exponential. To do so, we first note that the covariance of $(Y_1, Y_2, Y_3) := (X_1/\sqrt{p_1}, X_2/\sqrt{p_2}, X_3/\sqrt{p_3})$ is

$$\Sigma_Y = \begin{bmatrix} 1 - p_1 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} \\ -\sqrt{p_1p_2} & 1 - p_2 & -\sqrt{p_2p_3} \\ -\sqrt{p_1p_3} & -\sqrt{p_2p_3} & 1 - p_3 \end{bmatrix} = I - \begin{bmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{bmatrix} \times \begin{bmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{bmatrix}$$

Since $\mathbf{v}_1 = \begin{bmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{bmatrix}$ is a unit vector, we can complete it with two additional vectors $\mathbf{v}_2 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$ and $\mathbf{v}_3 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$ to form an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$. This can be done in many ways, for example by the Gram-Schmidt orthogonalization to $\mathbf{v}_1$, $[100]'$, $[010]'$. However, the specific form of $\mathbf{v}_2, \mathbf{v}_3$ does not enter the calculation - we only need to know that $\mathbf{v}_2, \mathbf{v}_3$ are orthonormal.

The point is that $I = \mathbf{v}_1 \mathbf{v}_1' + \mathbf{v}_2 \mathbf{v}_2' + \mathbf{v}_3 \mathbf{v}_3'$ as these are orthogonal eigenvectors of $I$ with $\lambda = 1$. (Or, because $\mathbf{x} = \mathbf{v}_1 \mathbf{v}_1' \mathbf{x} + \mathbf{v}_2 \mathbf{v}_2' \mathbf{v}_2 + \mathbf{v}_3 \mathbf{v}_3' \mathbf{x}$ as $\mathbf{v}_j' \mathbf{x} = \mathbf{x} \cdot \mathbf{v}_j$ are the coefficients of expansion of $\mathbf{x}$ in orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$.)

Therefore,

$$\Sigma_Y = \mathbf{v}_2 \mathbf{v}_2' + \mathbf{v}_3 \mathbf{v}_3'$$
We now notice that $\Sigma_Y$ is the covariance of the multivariate normal random variable $Z = v_2 Z_2 + v_3 Z_3$ where $Z_2, Z_3$ are independent real-valued $N(0, 1)$. Indeed,

$$EZZ' = \sum_{i,j=2}^{3} v_i v'_j E(Z_i Z_j) = \sum_{i=2}^{3} v_i v'_i$$

Therefore, vector $[Y_1 Y_2 Y_3]'$ has the same distribution as $Z$, and $Y_1^2 + Y_2^2 + Y_3^2$ has the same distribution as $\|Z\|^2 = \|v_2 Z_2 + v_3 Z_3\|^2 = \|v_2 Z_2\|^2 + \|v_3 Z_3\|^2 = Z_2^2 + Z_3^2$

(recall that $v_2$ and $v_3$ are orthogonal unit vectors).

\[\blacksquare\]

**Remark 12.17.** It is clear that this proof generalizes to all $k$.

We note that the distribution of $Z_2^2 + \cdots + Z_{k-1}^2$ is Gamma with parameters $\alpha = (k - 1)/2$ and $\beta = 2$, which is known under the name of chi-squared distribution with $k - 1$ degrees of freedom. To see that $Z_2^2 + Z_3^2$ is indeed chi-squared with two-degrees of freedom (i.e., exponential), we can determine the cumulative distribution function by computing $1 - F(u)$:

$$P(Z_2^2 + Z_3^2 > u) = \frac{1}{2\pi} \int_{x^2 + y^2 > u} e^{-(x^2+y^2)/2} dxdy = \frac{1}{2\pi} \int_0^{2\pi} \int_{r>\sqrt{u}} e^{-r^2/2} rdrd\theta = e^{-u/2}$$

### Required Exercises

**Exercise 12.1.** Suppose that $\mathbb{R}^{2k}$-valued random variables $(X_n, Y_n)$ are such that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 0$ (that is, $\lim_{n \to \infty} P(\|Y_n\| > \varepsilon) = 0$ for all $\varepsilon > 0$).

Prove that $X_n + Y_n \xrightarrow{D} X$.

**Exercise 12.2.** Suppose $(X_n, Y_n)$ are pairs of independent random variables and $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{D} Y$. Show that $(X_n, Y_n) \xrightarrow{D} \mu$ where $\mu$ is the product of the laws of $X$ and $Y$.

**Exercise 12.3.** Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables such that $E(\xi_1) = 0$, $E(\xi_1^2) = 1$. For $i = 1, 2, \ldots$, define $\mathbb{R}^2$-valued random variables $X_i = \begin{bmatrix} \xi_i \\ \xi_{i+1} \end{bmatrix}$ and let $S_n = \sum_{i=1}^{n} X_i$. Show that

$$\frac{1}{\sqrt{n}} S_n \xrightarrow{D} N(0, \Sigma)$$

for a suitable $2 \times 2$ covariance matrix $\Sigma$.

**Exercise 12.4.** Suppose $\xi_j, \eta_j, \gamma_j$ are i.i.d. mean zero variance 1. Construct the following vectors:

$$X_j = \begin{bmatrix} \xi_j - \eta_j \\ \eta_j - \gamma_j \\ \gamma_j - \xi_j \end{bmatrix}$$

Let $S_n = X_1 + \cdots + X_n$. Show that $\frac{1}{n} \|S_n\|^2 \xrightarrow{D} Y$. (In fact, $Y$ has gamma density.)
Chapter A

Review of math prerequisites

1. Convergence

1.1. Convergence of numbers. Recall that for a sequence of numbers, \( \lim_{n \to \infty} a_n = L \) means that ... (you should be able to write a formal definition!)

\[ \sum_{n=1}^{\infty} a_n = L \] means that ... (you should be able to write a formal definition!)

**Theorem A.1.** If a sequence of real numbers \( \{a_n\} \) is bounded and increasing, then \( \lim_n a_n = \sup_{n \in \mathbb{N}} a_n \).

For unbounded increasing sequences we write \( \lim_n a_n = \infty \).

Recall that for a sequence of numbers \( a_n \),

\[ \limsup_{n \to \infty} a_n = \lim_{n \to \infty} \sup_{k \geq n} a_k \quad \text{and} \quad \liminf_{n \to \infty} a_n = \lim_{n \to \infty} \inf_{k \geq n} a_k. \]

**Example A.1.** Here are a couple of examples of that illustrate several different possibilities, and also indicates standard conventions about the use of \( \infty \).

(i) Consider \( a_k = \frac{k}{k+1} \). Then \( \inf_{k \geq n} a_k = \min_{k \geq n} a_k = \frac{n}{n+1} \). So \( \liminf_{n \to \infty} a_n = 1 \). Similarly, \( \sup_{k \geq n} a_k = 1 \). So \( \limsup_{n \to \infty} a_n = 1 \).

(ii) Suppose \( a_k = (-1)^k \). Then \( \inf_{k \geq n} a_k = \min_{k \geq n} a_k = -1 \) and \( \sup_{k \geq n} a_k = \max_{k \geq n} a_k = 1 \). So \( \liminf_{n \to \infty} a_n = -1 \) and \( \limsup_{n \to \infty} a_n = 1 \).

(iii) Suppose \( a_k = (1 + (-1)^k) k \). Then \( \inf_{k \geq n} a_k = 0 \) and \( \sup_{k \geq n} a_k = \infty \). So \( \liminf_{n \to \infty} a_n = 0 \) and \( \limsup_{n \to \infty} a_n = \infty \).

**Remark A.2.** It is clear that \( \liminf_{n \to \infty} a_n \leq \limsup_{n \to \infty} a_n \). The equality holds iff the limit \( \lim_{n \to \infty} a_n \) exists as an extended number in \([-\infty, \infty]\).
Similarly, for a sequence of functions \( f_n : \Omega \rightarrow \mathbb{R} \), we define functions \( f_* , f^* : \Omega \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\} \) by \( f_* = \lim \inf_{n \rightarrow \infty} f_n \) and \( f^* = \lim \sup_{n \rightarrow \infty} f_n \) pointwise.

We say that the sequence of functions \( \{f_n \} \) converges pointwise, if \( f_n(\omega) \) converges for all \( \omega \in \Omega \).

We say that the sequence of functions \( \{f_n \} \) converges uniformly over \( \Omega \) to \( f \), if \( \sup_{\omega \in \Omega} |f_n(\omega) - f(\omega)| \rightarrow 0 \).

**Example A.2.** Consider \( f_n(x) = x^n \). If \( \Omega = [0, 1/2] \) then \( f_n \) converges uniformly over \( \Omega \). If \( \Omega = [0, 1] \) then \( f_n \) converges pointwise.

2. Set theory

(i) For a set \( \Omega \), by \( 2^\Omega \) we denote the so called power set, i.e., the set of all subsets of \( \Omega \). We use upper case letters like \( A,B,C, \ldots \) for the subsets - some (but not all) will be interpreted as "events".

(ii) The empty set is \( \emptyset \) - in handwriting this needs to be carefully distinguished from the Greek letters \( \varphi \) or \( \Phi \).

(iii) We use \( A \cup B \), for the union, \( A \cap B \) for the intersection, \( A^c \) or \( A' \) for the complement.

**We do not use** \( A + B \) and \( AB \) **in this course!!!**

(iv) We use \( A \subset B \) for what some other books denote by \( A \subseteq B \). Sometimes it will be convenient to write this as \( B \supset A \). Collections of sets will be denoted by scripted letters, like \( \mathcal{A} \) or \( \mathcal{F} \). We will need to consider large collections of sets, as well as collections like \( \mathcal{A} = \{A_1, A_2, \ldots \} \).

(v) For a family \( \mathcal{A} = \{A_t : t \in T\} \) of subsets of \( \Omega \) indexed by a set \( T \), the union of all sets in \( \mathcal{A} \) is the set of \( \omega \) with the property that there exists a set \( A_t \in \mathcal{A} \) such that \( \omega \in A_t \). In symbols,

\[
\bigcup_{t \in T} A_t = \{\omega \in \Omega : \omega \in A_t \text{ for some } t \in T \} = \{\omega \in \Omega : \exists t \in T \text{ } \omega \in A_t \}
\]

More concisely,

\[
\bigcup_{A \in \mathcal{A}} A = \{\omega \in \Omega : \exists A \in \mathcal{A} \text{ } \omega \in A \}
\]

Similarly, we define the intersection

\[
\bigcap_{t \in T} A_t = \{\omega \in \Omega : \forall t \in T \text{ } \omega \in A_t \}
\]

In particular, for a countable collection of sets,

\[
\bigcup_{n \in \mathbb{N}} A_n = \bigcup_{n=1}^{\infty} A_n = \{\omega : \omega \in A_n \text{ for some } n \in \mathbb{N} \}
\]

\[
\bigcap_{n \in \mathbb{N}} A_n = \bigcap_{n=1}^{\infty} A_n = \{\omega : \omega \in A_n \text{ for all } n \in \mathbb{N} \}
\]
(vi) The notation for intervals is \( (a, b) = \{ x \in \mathbb{R} : a < x < b \} \), \([a, b) = \{ x \in \mathbb{R} : a \leq x < b \}\) and similarly \((a, b]\) and \([a, b]\).

**Theorem A.3** (DeMorgan’s law).

\[
(A.1) \quad \left( \bigcup_{t \in T} A_t \right)^c = \bigcap_{t \in T} A_t^c
\]

Since \((A^c)^c = A\), formula (A.1) is equivalent to

\[
(A.2) \quad \left( \bigcap_{t \in T} A_t \right)^c = \bigcup_{t \in T} A_t^c
\]

**2.1. Indicator functions and limits of sets.** This has application to the so-called *indicator functions*:

\[
(A.3) \quad I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}
\]

Since

\[
I_{A_n}(\omega) = \begin{cases} 0 & \\ 1 \end{cases},
\]

it is clear that

\[
\limsup_{n \to \infty} I_{A_n}(\omega) = \begin{cases} 0 & \\ 1 \end{cases}.
\]

This means that \(\limsup_{n \to \infty} I_{A_n}(\omega) = I_{A^*}(\omega)\) for some set \(A^* \subset \Omega\).

For the same reasons, \(\liminf_{n \to \infty} I_{A_n} = I_{A_*}\) for some set \(A_* \subset \Omega\).

**Proposition A.4.**

\[
A_* = \bigcup_{n \in \mathbb{N}} \bigcap_{k \geq n} A_k \text{ and } A^* = \bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k
\]

**Proof.** This is Exercise A.1. \(\square\)

The second set has probabilistic interpretation:

\[
A^* = \{ A_n \text{ occur infinitely often } \} = \{ A_n \text{ i. o. } \}
\]

It is clear that \(A_* \subset A^*\). We say that \(\lim_{n \to \infty} A_n\) exists if \(A_* = A^*\). Exercises A.3 and A.4 give examples of such limits.

**2.2. Cardinality.** Sets \(A, B\) have the same *cardinality* if there exists a one-to-one and onto function \(f : A \to B\). We shall say that a set \(A\) is countable if either \(A\) is finite, or it has the same cardinality as the set \(\mathbb{N}\) of natural numbers.

It is known that the set of all rational numbers \(\mathbb{Q}\) is countable while the interval \([0, 1] \subset \mathbb{R}\) is not countable.
3. Compact set
Recall that if $K$ is compact if every sequence $\{x_n\}$ in $K$ has a convergent subsequence (with respect to some metric $d$). Equivalently, from every open cover of $K$ one can select a finite sub-cover.

If $K$ is compact and sets $F_n \subset K$ are closed with non-empty intersections $\bigcap_{k=1}^{n} F_k \neq \emptyset$ for all $n$, then the infinite intersection $\bigcap_{k=1}^{\infty} F_k$ is also non-empty.

**Theorem A.5.** Closed bounded subsets of $\mathbb{R}^k$ are compact.

**Proof.** To see what this involves, let's indicate why closed interval $[0, 1]$ is compact. \qed

4. Riemann integral
Function $f : [a, b] \to \mathbb{R}$ is Riemann-integrable, with integral $S = \int_{a}^{b} f(x)dx$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left| S - \sum_i f(x_j)|I_j| \right| < \varepsilon$$

for every partition of $[a, b]$ into sub-intervals $I_j$ of length $|I_j| < \delta$ and every choice of $x_j \in I_j$. Every Riemann-integrable function is Lebesgue-integrable over $[a, b]$.

It is known that continuous functions are Riemann-integrable.

In calculus, the improper integral $\int_0^{\infty} f(x)dx$ is defined as the limit $\lim_{t \to \infty} \int_0^{t} f(x)dx$. This is not the same as the Lebesgue integral over $[0, \infty)$.

5. Product spaces
The Cartesian product $U \times V$ of sets $U, V$ is the set of ordered pairs $\{(u, v) : u \in U, v \in V\}$. In particular, $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

The cartesian product $\times_{k=1}^{\infty} U_k$ is the set of all sequences $\{u_k\}$ of elements of the sets $U_1, U_2, \ldots$. In particular, $\mathbb{R}^{\infty} := \times_{k=1}^{\infty} \mathbb{R}$

The set $\mathbb{R}^{\infty}$ of all infinite sequences of real numbers is a metric space with the distance

$$d((a_n), (b_n)) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{|a_n - b_n|}{1 + |a_n - b_n|}.$$  

In particular, a sequence of points $a_n \in \mathbb{R}^{\infty}$ converges to $b$ if every coordinate converges. This is the pointwise convergence of functions, with a sequence $(a_n)$ identified with function $a : \mathbb{N} \to \mathbb{R}$.

6. Taylor polynomials and series expansions
See Chapter 12 Section 1 (page 131).

7. Complex numbers
See Chapter 12 Section 1 (page 131).
8. Metric spaces

Definition A.1. A function \(d : E \times E \rightarrow \mathbb{R}\) is called a metric if

(i) \(d(x, y) \geq 0\) and if \(d(x, y) = 0\) then \(x = y\)
(ii) \(d(x, y) = d(y, x)\)
(iii) \(d(x, z) \leq d(x, y) + d(y, z)\)

We then call the pair \((E, d)\) a metric space.

Here are some “elementary” examples of metric spaces.

(i) \(\mathbb{R}\) with \(d(x, y) = |x - y|\)
(ii) \(\mathbb{R}^d\) with \(d(x, y) = \|x - y\|\).
(iii) \([0, 1]\) with \(d(f, g) = \sup\{|f(x) - g(x)| : 0 \leq x \leq 1\}\).
(iv) The set of all CDFs on \(\mathbb{R}\) with Kolmogorov-Smirnov metric

\[
d(F, G) = \sup\{|F(x) - G(x)| : 0 < x < \infty\}
\]

(v) The set of all probability measures on \((\mathbb{R}, \mathcal{B})\) with total variation metric

\[
\delta(P, Q) = \sup\{|P(A) - Q(A)| : A \in \mathcal{B}\}
\]

It is known that \(\delta(P, Q) = \inf\{P(X \neq Y) : \mathcal{L}(X) = P, \mathcal{L}(Y) = Q\}\)

(vi) The set of all CDFs on \(\mathbb{R}\) with Levy’s metric:

\[
L(F, G) = \inf\{\varepsilon : G(x) \in [F(x - \varepsilon) - \varepsilon, F(x + \varepsilon) + \varepsilon]\text{ for all } x\}
\]

This is a metric for weak convergence: \(F_n \rightarrow F\) iff \(L(F_n, F) \rightarrow 0\).

Here are some of the metric spaces encountered in probability:

(i) The set of all CDFs (see Definition 2.3) on \(\mathbb{R}\) with Kolmogorov-Smirnov metric

\[
d(F, G) = \sup\{|F(x) - G(x)| : 0 < x < \infty\}
\]

This distance is used in statistics.

(ii) The set of all probability measures on \((\mathbb{R}, \mathcal{B})\) with total variation metric

\[
\delta(P, Q) = \sup\{|P(A) - Q(A)| : A \in \mathcal{B}\}
\]

It is known that \(\delta(P, Q) = \inf\{P(X \neq Y) : \mathcal{L}(X) = P, \mathcal{L}(Y) = Q\}\)

(iii) The set of all CDFs on \(\mathbb{R}\) with Levy’s metric:

\[
L(F, G) = \inf\{\varepsilon : G(x) \in [F(x - \varepsilon) - \varepsilon, F(x + \varepsilon) + \varepsilon]\text{ for all } x\}
\]

This is a metric for weak convergence: \(F_n \rightarrow F\) iff \(L(F_n, F) \rightarrow 0\), see Exercise Exercise 9.10.

(iv) The set of (classes of equivalence of) all random variables on \((\Omega, \mathcal{F}, P)\) with the distance

\[
d(X, Y) = E\left(\frac{|X - Y|}{1 + |X - Y|}\right)
\]

This is a metric for convergence in probability: \(X_n \xrightarrow{P} X\) iff \(d(X_n, X) \rightarrow 0\).
A. Review of math prerequisites

There are numerous other distances of interest. The following are frequently encountered and useful.

(i) The set of all (classes of equivalence of) integrable random variables $L_1(\Omega, \mathcal{F}, P)$ with the $L_1$ metric $\|X - Y\|_1 = E(|X - Y|)$

(ii) The set of all (classes of equivalence of) square integrable random variables $L_2(\Omega, \mathcal{F}, P)$ with the $L_2$ metric $\|X - Y\|_2 = \sqrt{E((X - Y)^2)}$.

(iii) The set of probability measures (CDFs) on $\mathbb{R}$ with the Waserstein distance

$$d(P, Q) = \inf \{ E|X - Y| : \mathcal{L}(X) = P, \mathcal{L}(Y) = Q \}$$

It is known that $d(P, Q) = \sup \int f(x)dP - \int f(x)dQ : f \text{Lipschitz with constant 1}$

---

**Additional Exercises**

**Exercise A.1.** Prove Proposition A.4.

**Exercise A.2.** Suppose $B, C$ are subsets of $\Omega$ and

$$A_n = \begin{cases} B & \text{if } n \text{ is even} \\ C & \text{if } n \text{ is odd} \end{cases}$$

Identify the sets $A_* = \bigcup_{n \in \mathbb{N}} \bigcap_{k \ge n} A_k$ and $A^* = \bigcap_{n \in \mathbb{N}} \bigcup_{k \ge n} A_k$.

**Exercise A.3.** Suppose $A_1 \supset A_2 \supset A_n \supset \ldots$. Show that $\lim_n A_n$ exists (and describe the limit).

**Exercise A.4.** Suppose $A_1 \subset A_2 \subset A_n \subset \ldots$. Show that $\lim_n A_n$ exists (and describe the limit).

**Exercise A.5.** If $a_n \to L$ for a finite $L$, show that $\{a_n\}$ is bounded.
Addenda

1. Modeling an infinite number of tosses of a coin

Consider a probability space \( \Omega = (0, 1] \) with field \( B_0 \) and probability measure \( \lambda \) as defined in Theorem 1.5. For \( \omega \in \Omega \), write \( \omega = \sum_{n=1}^{\infty} d_n(\omega)/2^n \) as the (non-terminating) binary expansion. For example, \( \omega = 1/2 \) has two such expansions:

\[
1/2 = 0/2^1 + 1/2^2 + 0/2^3 \cdots = 0.1 \quad \text{and} \quad 1/2 = 0/2 + 1/2^2 + 1/2^3 + \cdots = .0111\ldots
\]

and we take the second expansion, so that \( d_1(\omega) = 0 \) and \( d_2(\omega) = d_3(\omega) = \cdots = 1 \).

**Proposition B.1.** For a sequence of digits \( \varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\} \), the set

\[
A = \{ \omega : d_1(\omega) = \varepsilon_1, d_2(\omega) = \varepsilon_2, \ldots, d_n(\omega) = \varepsilon_n \}
\]

is in \( B_0 \) and has measure \( \lambda(A) = 1/2^n \).

**Proof.** We check that \( A = (\frac{k}{2^n}, \frac{k+1}{2^n}] \) is a diadic interval with \( k = 2^{n-1}\varepsilon_1 + 2^{n-2}\varepsilon_2 + \cdots + 2\varepsilon_{n-1} + \varepsilon_n \).

(Omitted in 2019) (The details were done in class.)

Here is a probabilistic interpretation: Toss a coin repeatedly, and label the outcomes of the tosses as 0 or 1. Here 1 represents "a success" and 0 represents a "failure". Event \( A \) occurs if a sequence of tosses produces the prescribed sequence of \( \varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\} \) - the prescribed "pattern" of failures and successes.

The probability that event \( A \) occurs is \( 1/2^n \), and it is the same for each of the \( 2^n \) possible choices of \( \varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\} \). Note however that \( n \) is not fixed, so we have a model for an infinite number of tosses. This infinite sequence of "tosses" arises from a single use of "a random number generator" that gives us \( \omega \in (0, 1] \) "at random".

The following additional facts could be worked out for this model.
(i) One can compute the probability of exactly \( k \) "successes" in \( n \) tosses. For \( n \in \mathbb{N} \) and integer \( 0 \leq k \leq n \), the set \( B = \{ \omega: d_1(\omega) + \cdots + d_n(\omega) = k \} \) is in \( \mathcal{B}_0 \) and \( \lambda(B) = \binom{n}{k}/2^n \) gives the probability for \( k \) "successes" in \( n \) tosses of a coin.

(ii) One can compute the probability that the \( n \)-th toss was successful. The set \( D_n = \{ \omega: d_n(\omega) = 1 \} \) is in \( \mathcal{B}_0 \) and we have \( \lambda(D_n) = 1/2 \).

(iii) The set \( C_1 = \{ \omega: d_n(\omega) = 1 \text{ for some } n \} = \bigcup_n D_n \) does not "have to be" in \( \mathcal{B}_0 \), but in fact it is (what is it?). Relying on our lifetime of experience with tossing coins, and based on the "probabilistic interpretation" that event \( C_1 \) occurs if we are successful at least once while tossing a coin repeatedly, we anticipate that we are "guaranteed" that \( C_1 \) occurs, that is we anticipate that \( \lambda(C) = 1 \).

(iv) The set \( D^* = \limsup_n D_n \) is a subset of \( C_1 \) which is also in \( \mathcal{B}_0 \). Since \( D^* = \{ D_n \ i.o. \} \) occurs when an infinite sequence of tosses results in an infinite number of successes, our experience with tossing the coins tells us that we should assign here probability \( \lambda(D^*) = 1 \), too.

(v) The set \( C = \{ \omega: d_n(\omega) = 0 \text{ for some } n \} = \bigcup_n D_n^c \) is not in \( \mathcal{B}_0 \), so we cannot compute its probability, yet. But of course it is in \( \sigma(\mathcal{B}_0) \), and precisely this kind of situation is the topic for the next set of lectures. Based on our lifetime of experience with tossing coins, and based on the "probabilistic interpretation" that event \( C \) occurs if we "fail" at least once while tossing a coin repeatedly, we would like to "improve" our model to ensure that we are "guaranteed" that \( C \) occurs, that is we would like to have \( \lambda(C) = 1 \).

(vi) The set \( F = \limsup_n D_n^c = (\liminf_n D_n)^c \) is a subset of \( C \) which is not in \( \mathcal{B}_0 \) either, so we cannot compute its probability, yet. But of course \( F \) is in \( \sigma(\mathcal{B}_0) \), and precisely this kind of situation is the topic for the next set of lectures. Since \( F = \{ D_n^c \ i.o. \} \) occurs when an infinite sequence of tosses results in an infinite number of failures, our experience with tossing the coins tells us that we should assign here probability \( \lambda(F) = 1 \), too.
Bibliography

[Gut] A. Gut, Probability: a graduate course
Index

$L_1$ metric, 11  
$L_2$ metric, 11  
$L_p$-norm, 59, 76  
$\lambda$-system, 25  
$\pi$-system, 25  
$\sigma$-field, 16  
$\sigma$-field generated by $X$, 41  

distribution of a random variable, 42  

Bernoulli random variables, 46  
Binomial distribution, 17, 73  
bivariate cumulative distribution function, 30  
Bonferroni’s correction, 18  
Boole’s inequality, 18  
Borel $\sigma$-field, 41  
Borel sigma-field, 16  

Cantelli’s inequality, 65  
cardinality, 9  
Cauchy distribution, 114  
Cauchy-Schwarz inequality, 58  
centered, 63  
Central Limit Theorem, 117  
characteristic function, 111  
characteristic function – continuity theorem, 114  
Characteristic functions – uniqueness, 113  
Characteristic functions – inversion formula, 113  
Chebyshev’s inequality, 57  
complex numbers, 108  
conjugate exponents, 60  
continuity condition, 14  
Continuous mapping theorem, 103  
converge in $L_p$, 61  
converge in mean square, 61  
convergence in distribution, 49, 99  
Convergence of types, 119  
converges in distribution, 127  
converges in probability, 46  
converges pointwise, 7  
converges uniformly, 7  
converges with probability one, 47  
convex function, 58  
correlation coefficient, 59  

countable additivity, 14  
covariance matrix, 130  
cumulative distribution function, 26, 43  
cylindrical sets, 32, 33  
cylindrical sets, 32  

de Moivre formula, 108  
DeMorgan’s law, 8  
density function, 29, 74  
diadic interval, 135  
discrete random variable, 73  
discrete random variables, 45  

equal in distribution, 43  
events, 13, 17  
expected value, 54, 69  
Exponential distribution, 74  
exponential distribution, 29  

Fatou’s lemma, 71  
field, 13  
finitely dimensional distributions, 32  
finitely-additive probability measure, 14  
Fubini’s Theorem, 82  

Geometric distribution, 73  
Hölder’s inequality, 60, 76  

inclusion-exclusion, 18  
independent $\sigma$-fields, 35  
independent events, 35  
independent identically distributed, 46  
independent random variables, 44  
indicator functions, 9  
induced measure, 42  
infinite number of tosses of a coin, 135  
integrable, 69  
intersection, 8  

Jensen’s inequality, 58  
joint cumulative distribution function, 30  
joint distribution of random variables, 43
Kolmogorov’s maximal inequality, 91
Kolmogorov’s one series theorem, 91
Kolmogorov’s three series theorem, 92
Kolmogorov’s two series theorem, 92
Kolmogorov’s zero-one law, 90
Kolmogorov-Smirnov metric, 10, 11
Kronecker’s Lemma, 94
Lebesgue’s dominated convergence theorem, 71, 72
Lebesgue’s dominated convergence theorem – used, 73, 89, 102, 115
Levy’s metric, 11
Levy’s theorem, 93
Lindeberg condition, 120
Lyapunov’s condition, 121
Lyapunov’s inequality, 58
marginal cumulative distribution functions, 30
Markov’s inequality, 57
maximal inequality, Etemadi’s, 95
maximal inequality, Kolmogorov’s, 91
mean square convergence, 76
measurable function, 41
measurable rectangle, 81
metric, 10
metric space, 10
Minkowski’s inequality, 59
Minkowski’s inequality, 76
moment generating function, 56, 77
moments, 55
Monotone Convergence Theorem, 69
multivariate normal, 129
multivariate normal distribution, 130
multivariate random variable, 41
negative binomial distribution, 18
normal distribution, 29
Poisson distribution, 18, 73
Polya’s distribution, 18
Portmanteau Theorem, 102
power set, 7
probability, 13
probability measure, 14
probability space, 13, 17
product measure, 82
quantile function, 44, 101
random element, 41
random variable, 41
random vector, 41
sample space, 13
Scheffe’s theorem, 100
section, 81
semi-algebra, 15
semi-ring, 15
sigma-field generated by \( \mathcal{A} \), 16
simple random variable, 53
simple random variables, 45
Skorohod’s theorem, 101
Slutsky’s Theorem, 101
Standard normal density, 74
stochastic process with continuous trajectories, 43
stochastic processes, 42
stochastically bounded, 51
symmetric distribution, 97
tail \( \sigma \)-field, 36
Tail integration formula, 84
Taylor polynomials, 107
tight, 51
tight probability measure, 19
Tonelli’s theorem, 82
total variation metric, 11
truncation of r.v., 50
uncorrelated, 63
uniform continuous, 28
Uniform density, 74
uniform discrete, 28
uniform singular, 28
uniformly integrable, 72, 105
union, 8
variance, 56
Wasserstein distance, 11
weak convergence, 49
weak law of large numbers, 63
zero-one law, 36, 90