Chapter H

A Primer on Probability Limit Theorems

We now know that the independence of a sequence of random variables does not restrict what sort of a distribution any one of these random variables may have. But this does not mean that independence does not have deep consequences about the behavior of such a sequence. Indeed, various types of random sequences made up of countably many independent random variables possess distinguishing types of limiting behavior. A major theme in probability theory is the determination of the asymptotic behavior of such sequences. This chapter is devoted to this theme, but our introduction is only intended to be a brief, yet hopefully appetizing, primer.

We begin the chapter with introducing an additional convergence concept for random variables, namely, convergence in probability. After examining how this notion of convergence relate to other convergence notions we have encountered so far in the text, we move on to study the laws of large numbers. We examine various forms of the Weak and Strong Law of Large Numbers, and go through some applications of these laws to inferential statistics. We then move to studying some important zero-one laws of probability limit theory. Finally, in a section that is slightly more advanced than the rest, we study the convergence of series of independent random variables, and, in particular, establish Lévy’s Theorem and the Kolmogorov Convergence Criterion.¹

1 Preliminaries

To study the limit behavior of a sequence (or series) of independent random variables, we must, of course, first have to agree on what we mean by “limit” here. After all, probability theory has quite a number different such notions, each being useful towards different ends. We have already encountered two distinct convergence concepts, namely, weak limit and almost sure limit of a random sequence. There are other useful modes of convergence in probability theory. In particular, essential for our present study is the one called probabilistic limit. After revisiting the notion of almost sure convergence, we shall thus begin with a thorough discussion of this convergence concept as a first step in our asymptotic probability analysis.

¹Most of the topics covered in this chapter would be covered in any graduate text on probability theory. Each of the references mentioned in Chapter B, for instance, provide more advanced (and complete) treatments of the probability limit theorems. But let me note that statistically oriented probability texts, such as Chow and Teicher (1997) and Gut (2005), go into this topic more deeply than others, paying due attention to issues related to rate of convergence, among others.
1.1 Upper and Lower Limits of Events

First Impressions

As you have surely noticed by now, monotonic sequences of events figure quite frequently in probability theory. This is mainly because it is particularly easy to study the probabilistic behavior of such a sequence in the limit. Be that as it may, even when a sequence of events is not monotonic, we may still talk about its limiting behavior. The idea is analogous to the use of the upper and lower limits of a non-convergent real sequence in order to gather asymptotic information about it.

**Definition.** Let $X$ be a nonempty set and $S_m \subseteq X$, $m = 1, 2, \ldots$. We define
\[
\limsup S_m := \bigcap_{k=1}^{\infty} \bigcup_{i=k}^{\infty} S_i \quad \text{and} \quad \liminf S_m := \bigcup_{k=1}^{\infty} \bigcap_{i=k}^{\infty} S_i.
\]

$\limsup S_m$ is called the **upper limit** of the sequence $(S_m)$, and $\liminf S_m$ is called its **lower limit**. If $\limsup S_m = \liminf S_m$, then we say that the sequence $(S_m)$ is **convergent** and write $\lim S_m$ for the common value of $\limsup S_m$ and $\liminf S_m$.

A moment’s reflection shows that if $(S_m)$ is a sequence of subsets of a nonempty set $X$, then $\limsup S_m$ is the set of all elements of $X$ that belong to infinitely many terms of the sequence. That is,
\[
\limsup S_m = \{\omega \in X : \omega \in S_m \text{ for infinitely many } m\}.
\]
(This is important – please think about this equation until it becomes trivial to you.) Similarly, we have
\[
\liminf S_m = \{\omega \in X : \omega \in S_m \text{ for all but finitely many } m\}.
\]
(Why?) In probabilistic jargon, one often writes $\limsup S_m = \{S_m \text{ infinitely often}\}$ and $\liminf S_m = \{S_m \text{ eventually}\}$. At times we will adopt this convention as well.

The following exercise collects together some useful properties of the upper and lower limits of a sequence of sets. We will use these properties freely in what follows.

**Exercise 1.1.** Let $(S_m)$ be a sequence of subsets of a nonempty set $X$. Prove:
(a) $\liminf S_m \subseteq \limsup S_m$,
(b) $X \setminus \limsup S_m = \liminf X \setminus S_m$,
(c) $\limsup 1_{S_m} = 1_{\limsup S_m}$ and $\liminf 1_{S_m} = 1_{\liminf S_m}$,
(d) If $(S_m)$ is increasing (or decreasing), then it is convergent.

Another fundamental fact to keep in mind is that the limsup and liminf of a sequence of measurable sets (in a measurable space) are themselves measurable. That is, if $S_1, S_2, \ldots$ belong to a $\sigma$-algebra $\Sigma$ on a nonempty set $X$, then both $\limsup S_m$
and \( \lim \inf S_m \) also belong to \( \Sigma \), simply because a \( \sigma \)-algebra is closed under taking countable unions and intersections.

**Insight.** The upper and lower limits of any sequence of events (in a probability space) can always be assigned probability values.

**A Fatou Lemma for Limits of Events**

A natural question is, then, how the probabilities of the upper and lower limits of a sequence of events relate to the probabilities of the terms of that sequence. A basic result in this regard is the following by-product of Fatou’s Lemma.

**Lemma 1.1.** Let \((X,\Sigma,p)\) be a probability space and \((S_m)\) a sequence in \(\Sigma\). Then,

\[
p(\lim \inf S_m) \leq \lim \inf p(S_m) \leq \lim \sup p(S_m) \leq p(\lim \sup S_m).
\]

**Proof.** As we have noted in Exercise 1.1, \(\lim \inf 1_{S_m} = 1_{\lim \inf S_m}\). Thus, by Fatou’s Lemma,

\[
p(\lim \inf S_m) = \int_X 1_{\lim \inf S_m} \, dp = \int_X \lim \inf 1_{S_m} \, dp \leq \lim \inf \int_X 1_{S_m} \, dp,
\]

that is, \(p(\lim \inf S_m) \leq \lim \inf p(S_m)\). The second inequality in the claim is trivial, and the third is deduced from the first by using the fact that \(X \setminus \lim \sup S_m = \lim \inf X \setminus S_m\).

Exercise 1.2. Give examples to show that either of the inequalities in Lemma 1.1 may hold strictly.

**1.2 Almost Sure Convergence, Revisited**

Recall that a sequence \((x_m)\) of random variables on a probability space \((X,\Sigma,p)\) almost surely converges to some \(x \in \mathcal{L}^0(X,\Sigma)\) — that is, \(x_m \to_{\text{a.s.}} x\) — if the probability of the event that \(x_m\) does not converge to \(x\) (pointwise) is zero, or put succinctly, 

\[p\{x_m \to x\} = 1\].

This is a pleasant formulation that makes it clear why we would be interested in such a notion of convergence. However, in practice, it is difficult to use this definition to verify that \(x_m \to_{\text{a.s.}} x\) directly. Instead, it is often more convenient to work with an alternative formulation of almost sure convergence, one that is based on upper limits of events. We give this formulation next.

**Lemma 1.2.** Let \(Y\) be a separable metric space, and \(x, x_1, x_2, \ldots\) \(Y\)-valued random variables on a probability space \((X,\Sigma,p)\). Then, \(x_m \to_{\text{a.s.}} x\) if, and only if,

\[
p(\lim \sup \{d_Y(x_m, x) > \varepsilon\}) = 0 \quad \text{for every} \quad \varepsilon > 0.
\]

\[\text{Reminder.} \quad \text{The metric of} \ Y \ \text{is denoted as} \ d_Y.\]
Proof. The “only if” part of this assertion is fairly obvious, so we focus only on its “if” part. The idea is to use (1) for arbitrarily small \( \varepsilon > 0 \). To this end, let us assume (1), and define
\[
S_k := \limsup \left\{ d_Y(x_m, x) > \frac{1}{k} \right\}, \quad k = 1, 2, \ldots
\]
Since \( S_1 \subseteq S_2 \subseteq \cdots \), we have \( S_k \not= \bigcup^n \infty S_i =: S \). By (1), we have \( p(S_1) = p(S_2) = \cdots = 0 \), so, by the continuity of probability measures, \( p(S) = 0 \), that is, \( p(X \setminus S) = 1 \). But \( \{x_m \to x\} \) contains \( X \setminus S \). To see this, take any \( \omega \in X \setminus S \). Then, \( \omega \) is not in \( S_k \) for any positive integer \( k \). That is, for every \( k \in \mathbb{N} \), we have \( d_Y(x_m(\omega), x(\omega)) \leq \frac{1}{k} \) for all but finitely many \( m \), which means \( x_m(\omega) \to x(\omega) \).

1.3 Convergence in Probability

First Impressions

The notion of almost sure convergence often turns out to be too demanding for the analysis of the asymptotic behavior of a random sequence. In such situations, one needs a somewhat weaker mode of convergence. There are several intriguing alternatives in this regard, but one that is particularly useful is the notion of convergence in probability.

Definition. Let \( Y \) be a separable metric space, and \( x, x_1, x_2, \ldots \) \( Y \)-valued random variables on a probability space \((X, \Sigma, p)\). We say that \((x_m)\) \textbf{converges to} \( x \) \textbf{in probability}, and write \( x_m \to x \) in probability, or \( p\lim x_m = x \), if
\[
p\{d_Y(x_m, x) > \varepsilon\} \to 0 \quad \text{for every } \varepsilon > 0.
\]
That is, \( x_m \to x \) in probability iff, for every positive real numbers \( \varepsilon \) and \( \delta \), there exists a positive integer \( M \) such that
\[
p\{d_Y(x_m, x) > \varepsilon\} < \delta \quad \text{for all } m \geq M.
\]

In other words, for any separable metric space \( Y \), a sequence of \( Y \)-valued random variables converges to a \( Y \)-valued random variable \( x \) in probability if all of these random variables are defined on the same probability space – provided that the probability that the sequence will approximate \( x \) to any desired degree of accuracy is arbitrarily close to 1. (Here, of course, “approximation” is relative to how “distance” is measured in \( Y \).) In particular, a sequence \((x_m)\) of random variables converges to a random variable \( x \) in probability iff the sequence \( (p\{ |x_m - x| > \varepsilon \}) \) vanishes in the limit no matter how small \( \varepsilon \) is.

\(^3\)Since \( d_Y \in C(Y \times Y) \) and \( Y \) is separable, \( d_Y(x_m, x) \) is a random variable on \((X, \Sigma, p)\). (Recall Example B.5.4.) Consequently, \( \{d_Y(x_m, x) > \varepsilon\} \) belongs to \( \Sigma \) for any \( \varepsilon > 0 \), and hence the notion of convergence in probability is well-defined for random variables that take values in a separable metric space.
Let’s first see how this convergence concept relates to the previous two modes of convergence that we have encountered in this course. Things are fairly straightforward with respect to convergence in distribution (Section D.2.1).

**Convergence in Probability implies Convergence in Distribution**

**Proposition 1.1.** Let \( Y \) be a separable metric space and \( x, x_1, x_2, \ldots \) \(-\)valued random variables on a probability space \((X, \Sigma, p)\). If \( x_m \to x \) in probability, then \( x_m \overset{D}{\to} x \).

**Proof.** By Corollary D.2.2, it is enough to show that \( \text{p-lim } x_m = x \) implies \( \mathbb{E}(g \circ x_m) \to \mathbb{E}(g \circ x) \) for every bounded and Lipschitz continuous real map \( g \) on \( Y \). Let us then fix an arbitrary \( g \in \mathcal{B}(Y) \) such that there exists a real number \( K > 0 \) with \( |g(\nu_1) - g(\nu_2)| \leq Kd_Y(\nu_1, \nu_2) \) for every \( \nu_1, \nu_2 \in Y \). Take any \( \varepsilon > 0 \), and define \( S_m := \{ d_Y(x_m, x) > K\varepsilon \} \) for every positive integer \( m \). Then, \( \text{p-lim } x_m = x \) implies that there exists a positive integer \( M \) large enough that \( \text{p}(S_m) = 0 \), and hence \( \mathbb{E}(d_Y(x_m, x)1_{S_m}) = 0 \) for each \( m \geq M \). (Yes?) Consequently,

\[
|\mathbb{E}(g \circ x_m) - \mathbb{E}(g \circ x)| \leq \int_X |g \circ x_m - g \circ x| \, dp \\
\leq K \left( \int_{X \setminus S_m} d_Y(x_m, x) \, dp + \int_{S_m} d_Y(x_m, x) \, dp \right) \\
\leq K \frac{\varepsilon}{K} = \varepsilon
\]

for each \( m \geq M \). Conclusion: \( \mathbb{E}(g \circ x_m) \to \mathbb{E}(g \circ x) \).

It is easy to see that the converse of Proposition 1.1 is false in general.

**Example 1.1.** Take the probability space \((\{0, 1\}, 2^{\{0,1\}}, p)\) where both \( p\{0\} \) and \( p\{1\} \) are equal to \( \frac{1}{2} \), and consider the following random variables defined on this space:

\[
x(\omega) := \begin{cases} 
1, & \text{if } \omega = 0 \\
0, & \text{if } \omega = 1 
\end{cases} \quad \text{and} \quad x_m(\omega) := \begin{cases} 
0, & \text{if } \omega = 0 \\
1, & \text{if } \omega = 1 
\end{cases}
\]

for each positive integer \( m \). We obviously have \( x_m \overset{D}{\to} x \), because the distribution functions of each of these random variables are identical. Yet, clearly, \( (x_m) \) does not converge to \( x \) in probability.

**Almost Sure Convergence implies Convergence in Probability**

The following result shows that convergence in probability sits, in general, between almost sure convergence and convergence in distribution.
Proposition 1.2. (Kolmogorov) Let \( Y \) be a separable metric space and \( x, x_1, x_2, \ldots \) \(-\)valued random variables on a probability space \((X, \Sigma, \mathbb{P})\). If \( x_m \to_{\text{a.s.}} x \), then \( x_m \to x \) in probability.\(^4\)

**Proof.** Take any \( \varepsilon > 0 \), and set \( S_m := \{d_Y(x_m, x) > \varepsilon\} \) for each positive integer \( m \). By Lemmas 1.1 and 1.2, \( x_m \to_{\text{a.s.}} x \) implies \( \limsup \mathbb{P}(S_m) \leq \mathbb{P}(\limsup S_m) = 0 \). As \( \mathbb{P}(S_m) \geq 0 \) for each \( m \), it follows that \( x_m \to_{\text{a.s.}} x \) implies \( \mathbb{P}(S_m) \to 0 \).

The converse of this result is also false, as we show next.

Example 1.2. Consider the sequence \( (x_m) \) of random variables on \(([0, 1], \mathcal{B}[0, 1], \ell)\) defined by \( x_1 := \text{id}_{[0,1]}, x_2 := x_1 + 1_{[0,1/2]}, x_3 := x_1 + 1_{[1/2,1]}, x_4 := x_1 + 1_{[0,1/3]}, x_5 := x_1 + 1_{[1/3,2/3]}, x_6 := x_1 + 1_{[2/3,1]}, \) and so on. Clearly, let alone \( (x_m) \) being almost surely convergent, \( (x_m(\omega)) \) is not convergent for any \( \omega \in [0, 1] \) in this example. Yet,

\[
(\ell\{|x_m - x_1| > \varepsilon\}) = (0, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \ldots)
\]

for every \( \varepsilon > 0 \), that is, \( x_m \to x_1 \) in probability.

**Insight.**

| almost sure convergence | \( \implies \) | convergence in probability | \( \implies \) | convergence in distribution |

and the converse of any one of these implications is false.

**Exercise 1.3.** Let \( (x_m) \) be a sequence of independent random variables on a probability space \((X, \Sigma, \mathbb{P})\) such that \( \mathbb{P}\{x_m = 1\} = \frac{1}{m} = 1 - \mathbb{P}\{x_m = 0\} \) for each \( m \). Show that \( \mathbb{P}\text{-}\lim x_m = 0 \) but \( x_m \to_{\text{a.s.}} 0 \) is false.

**Exercise 1.4.** Let \( (x_m) \) be a sequence of independent random variables on a probability space \((X, \Sigma, \mathbb{P})\) such that \( \mathbb{P}\{x_m = 1\} + \mathbb{P}\{x_m = 0\} = 1 \) for each \( m \). Prove:

(a) \( \mathbb{P}\text{-}\lim x_m = 0 \) iff \( \lim \mathbb{P}\{x_m = 1\} = 0 \);

(b) \( x_m \to_{\text{a.s.}} 0 \) iff \( \sum_{\infty} \mathbb{P}\{x_i = 1\} < \infty \).

**Exercise 1.5.** Let \( (x_m) \) be a sequence of independent random variables on a probability space \((X, \Sigma, \mathbb{P})\). Show that if \( \mathbb{P}\text{-}\lim x_m = x \) for some \( x \in L^0(X, \Sigma) \), then \( x \) must be almost surely constant.

\(^4\) **Warning.** The notion of convergence in probability extends, in the obvious way, to measurable functions defined on an arbitrary measure space; in such a context, it is called *convergence in measure*. In the case of an arbitrary finite measure space, a.s. convergence is stronger than convergence in measure – the proof of this is analogous to the one I’m about to give. However, in the case of infinite measure spaces, a.s. convergence does not imply convergence in measure. (Quiz. Try giving an example that shows this. **Hint.** The second part of Proposition B.2.2 fails in infinite measure spaces.)
Exercise 1.6. Let $Y$ and $Z$ be two separable metric spaces. Let $(x_m)$ be a sequence of $Y$-valued random variables on a probability space that converges to a $Y$-valued random variable $x$ in probability. Show that $f(x_m) \to f(x)$ in probability for any continuous function $f : Y \to Z$.

Exercise 1.7. Let $(x_m)$ and $(y_m)$ be two sequences of random variables on a probability space $(X, \Sigma, p)$. Suppose that $x_m \overset{D}{\to} x$ for some $x \in \mathcal{L}^0(X, \Sigma)$, while $p$-lim $y_m = 0$. Prove: $x_m + y_m \overset{D}{\to} x$.

Exercise 1.8. (The Cauchy Criterion) Let $Y$ be a Polish space and $(x_m)$ a sequence of $Y$-valued random variables on a probability space $(X, \Sigma, p)$. Prove that $(x_m)$ converges in probability iff for every $\varepsilon, \delta > 0$ there is a positive integer $M$ such that

$$p\{d_Y(x_k, x_l) > \varepsilon\} < \delta \quad \text{for all } k, l \geq M.$$

Exercise 1.9. Let $(x_m)$ be a sequence of nonnegative random variables on a probability space $(X, \Sigma, p)$. Prove that

$$p\text{-lim } x_m = 0 \quad \text{iff} \quad \mathbb{E}\left(\frac{x_m}{1 + x_m}\right) \to 0.$$

Exercise 1.10. Let $(x_m)$ be a sequence of random variables on a probability space $(X, \Sigma, p)$, and let $x \in \mathcal{L}^0(X, \Sigma)$. Show that if $X$ is countable and $p$-lim $x_m = x$, then $x_m \to_{a.s.} x$.

Exercise 1.11. Let $Y$ be a separable metric space, and $\mathcal{X}$ stand for the set of all $Y$-valued random variables on a probability space $(X, \Sigma, p)$. Show that the Ky Fan metric $d^F$ on $\mathcal{X}$ introduced in Exercise E.3.8 metrizes convergence in probability (that is $p$-lim $x_m = x$ iff $d^F(x_m, x) \to 0$ for every $x, x_1, x_2, \ldots$ in $\mathcal{X}$).

2 Laws of Large Numbers

2.1 Weak Law of Large Numbers

The Classical Weak Law

Consider a situation in which a given experiment is to be repeated an indefinite number of times, and we are interested in a particular statistic that will arise from these experiments on average. (For instance, it would be nice if we could say something intelligent about the average earnings of an investor who invests a fixed amount on a particular risky prospect over and over again.) To study this sort of a situation in the abstract, we would take a sequence $(x_m)$ of independently and identically distributed random variables, and investigate the asymptotic behavior of the random sequence $(\frac{1}{m}(x_1 + \cdots + x_m))$. As the values of the $x_m$s are drawn independently according to a fixed probability distribution, it seems plausible that the sample average $\frac{1}{m}(x_1 + \cdots + x_m)$ (which is random) would then cumulate around the population average $\mathbb{E}(x_1)$ (which is not random).

There are various theorems in probability theory which formalize this intuition—such results often bear the name “laws of large numbers.” A very first such theorem was proved by Jacob Bernoulli in 1712 (in the context of sequences of independent binary random variables). While Bernoulli’s argument was quite involved, there have
appeared in time numerous generalizations of his law of large numbers, often with much simpler proofs. Among these, the following – established by Pafnuty Chebyshev in 1867 – is one of the most important.

**The Weak Law of Large Numbers.** (Chebyshev) Let \((x_m)\) be a sequence of independent random variables on a probability space \((X, \Sigma, p)\) with \(\mathbb{E}(x_1) = \mathbb{E}(x_2) = \cdots \in \mathbb{R}\) and \(\sup \mathbb{V}(x_m) < \infty\). Then,

\[
\mathbb{E}\left(\left|\frac{1}{m} \sum_{i \in [m]} x_i - \mathbb{E}(x_1)\right|\right) \to 0 \quad \text{and} \quad \frac{1}{m} \sum_{i \in [m]} x_i \to \mathbb{E}(x_1) \quad \text{in probability.}
\]

**Proof.** Let \(\mu := \mathbb{E}(x_1)\), \(s := \sup \mathbb{V}(x_m)\) and define \(y_m := \frac{1}{m}(x_1 + \cdots + x_m)\) for each positive integer \(m\). Then, \(\mathbb{E}(y_m) = \mu\), and by Exercise G.2.13, \(\mathbb{V}(y_m) = \frac{1}{m^2}(\mathbb{V}(x_1) + \cdots + \mathbb{V}(x_m)) \leq \frac{s}{m}\) for each \(m\). Therefore, by Jensen’s Inequality,

\[
\mathbb{E}(|y_m - \mu|) \leq \sqrt{\mathbb{E}((y_m - \mu)^2)} = \sqrt{\mathbb{V}(y_m)} \to 0
\]

as \(m \to \infty\). Our second assertion follows from the first by means of Markov’s Inequality.

The following is a special case of the Weak Law of Large Numbers that is worth stating separately. It is sometimes referred to as the Classical Weak Law of Large Numbers in the literature.

**Corollary 2.1.** Let \((x_m)\) be a sequence of i.i.d. random variables with finite expectation and variance. Then,

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to \mathbb{E}(x_1) \quad \text{in probability.}
\]

**Generalizations of the Classical Weak Law of Large Numbers**

Corollary 2.1 is actually not a first-best result. It turns out that by using a suitable truncation argument we can establish the same conclusion without assuming anything about the variances of the involved random variables. This result, which was established by Alexander Khinchine in 1928, is routinely utilized in statistics.\(^5\)

**Khinchine’s Weak Law of Large Numbers.** Let \((x_m)\) be a sequence of i.i.d. random variables with finite expectation. Then,

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to \mathbb{E}(x_1) \quad \text{in probability.}
\]

\(^5\) A statistician would read this as saying that the sample mean is a *consistent* estimator for the population mean (provided that sample selection is performed independently). See Section 2.5 for more on this.
Proof. Let \( \mu := \mathbb{E}(x_1) \), and define \( y_m := \frac{1}{m}(x_1 + \cdots + x_m) \) for each positive integer \( m \). Notice that \( \mathbb{E}(y_m) = \mu \) for each \( m \), and hence, thanks to Markov’s Inequality, it is enough to prove that \( \mathbb{E}(|y_m - \mu|) \to 0 \). To this end, fix a positive integer \( K \), and consider the truncated random variables \( x_{i,K} := x_i \mathbb{1}_{\{|x_i| \leq K\}} \) for each \( i \in \mathbb{N} \), which, obviously, have finite variance. Now define \( y_{m,K} := \frac{1}{m}(x_{1,K} + \cdots + x_{m,K}) \) for each \( m \). By the Triangle Inequality,
\[
\mathbb{E}(|y_m - \mu|) \leq \mathbb{E}(|y_m - y_{m,K}|) + \mathbb{E}(|y_{m,K} - \mathbb{E}(x_{1,K})|) + \mathbb{E}(|\mathbb{E}(x_{1,K}) - \mu|).
\]

It is easy to estimate the right-hand side of this inequality. Indeed, as the distributions of \( x_i \) are identical, we have
\[
\mathbb{E}(|y_{m,K} - \mathbb{E}(x_{1,K})|) = \frac{1}{m} \sum_{i \in [m]} \int |x_i - x_{i,K}| \, dp = \frac{1}{m} \sum_{i \in [m]} \int_{\{|x_i| > K\}} |x_i| \, dp = \int_{\{|x_1| > K\}} |x_1| \, dp.
\]
while
\[
\mathbb{E}(|\mathbb{E}(x_{1,K}) - \mu|) = |\mathbb{E}(x_{1,K}) - \mathbb{E}(x_1)| \leq \mathbb{E}(|x_{1,K} - x_1|) = \int_{\{|x_1| > K\}} |x_1| \, dp.
\]
Consequently,
\[
\mathbb{E}(|y_m - \mu|) \leq 2 \int_{\{|x_1| > K\}} |x_1| \, dp + \mathbb{E}(|y_{m,K} - \mathbb{E}(x_{1,K})|).
\]
As the Weak Law of Large Numbers ensures \( \mathbb{E}(|y_{m,K} - \mathbb{E}(x_{1,K})|) \to 0 \), therefore,
\[
\limsup \mathbb{E}(|y_m - \mu|) \leq 2 \int_{\{|x_1| > K\}} |x_1| \, dp.
\]
But we have established this for an arbitrary positive integer \( K \). Since
\[
\int_{\{|x_1| > K\}} |x_1| \, dp \to 0 \quad \text{as } K \to \infty,
\]
by the Dominated Convergence Theorem, we thus find \( \limsup \mathbb{E}(|y_m - \mu|) = 0 \), which means \( \mathbb{E}(|y_m - \mu|) \to 0 \), as we sought.

Another direction in which we can generalize the Weak Law of Large Numbers is by weakening its independence assumption. In particular, it is not difficult to show that this law applies to sequences of uncorrelated random variables.

Exercise 2.1.\(^H\) Show that the conclusion of the Weak Law of Large Numbers would remain unchanged, if we replaced the independence requirement in its statement with the hypothesis that \( \mathbb{E}(x_i x_j) = \mathbb{E}(x_i)\mathbb{E}(x_j) \) for every distinct positive integers \( i \) and \( j \).
In fact, we can say quite a bit more in this regard. In particular, it is possible to weaken the independence and same-means assumptions simultaneously in the statement of the Weak Law of Large Numbers. The following result is prototypical of this kind of generalizations. It was obtained by Andrei Markov in 1907.\footnote{Andrei Markov (1856-1922) was a gifted student of Chebyshev. His attempts on weakening the independence assumption in the Weak Law of Large Numbers have led him to the discovery of what we today call \textit{Markov chains}, and made Markov one of the founders of the theory of stochastic processes. If you want to learn more about the life and contributions of Markov, let me mention that Basharin, Langville and Naumov (2004) is a very enjoyable read.}

**Markov’s Weak Law of Large Numbers.** Let \((x_m)\) be a sequence of uncorrelated random variables on a probability space \((X, \Sigma, p)\) such that \(\sup \mathbb{E}(x_m) < \infty\) and \(\sup \mathbb{V}(x_m) < \infty\). Assume further that

\[
\lim \frac{1}{m} \sum_{i \in [m]} x_i \geq \mathbb{E}(x_i) < \infty \quad \text{and} \quad \frac{1}{m^2} \sum_{i \in [m]} \mathbb{V}(x_i) \to 0.
\]

Then,

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to \lim \frac{1}{m} \sum_{i \in [m]} \mathbb{E}(x_i) \quad \text{in probability.}
\]

We present the proof of this result in the form of an exercise.

**Exercise 2.2.** Let \((x_m)\) be as in the statement of Markov’s Weak Law of Large Numbers. Let \(\mu_m := \mathbb{E}(x_m)\) for each \(m\), and define \(\mu := \lim m^{-1} \sum_{i \in [m]} \mu_i\), which is well-defined by hypothesis.

(a) Define \(z_m := \sum_{i \in [m]} (x_i - \mu_i) / m\) for each positive integer \(m\), and show that \(\mathbb{E}(z_m) = 0\) and \(\mathbb{V}(z_m) = \sum_{i \in [m]} \mathbb{V}(x_i) / m^2\) for each \(m\).

(b) Use the Chebyshev-Bienaymé Inequality to show that \(p\)-\(\lim z_m = 0\).

(c) Take any \(\varepsilon > 0\), and show that

\[
\left\{ \left| \frac{1}{m} \sum_{i \in [m]} x_i - \mu \right| > \varepsilon \right\} \subseteq \{|z_m| > \frac{\varepsilon}{2}\},
\]

for all but finitely many \(m \in \mathbb{N}\), and combine this with part (b) to complete the proof of Markov’s Weak Law of Large Numbers.

**Exercise 2.3.** (Bernstein) Show that, in the statement of Markov’s Weak Law of Large Numbers, one can replace the second assumption in (2) with the following: There exist a number \(K > 0\) and a real sequence \((a_m)\) such that

\[
\text{(i)} \quad \sum_{i \in [m]} \mathbb{V}(x_i) < Km, \quad m = 1, 2, \ldots; \quad \text{and}
\]

\[
\text{(ii)} \quad \frac{1}{m} \sum_{i \in [m]} a_i \to 0 \quad \text{and} \quad \text{Cor}(x_i, x_j) \leq a_{|i-j|} \quad \text{for any distinct positive integers } i \text{ and } j.
\]
2.2 Application: The Weierstrass Approximation Theorem

The laws of large numbers have many interesting applications, and surprisingly, some of these are not probabilistic in spirit. In particular, the Weak Law of Large Numbers provides us with a general method for constructing a sequence of non-degenerate probability measures that approximate a degenerate random variable. A smart choice of such a sequence may then enable us to convert a non-probabilistic problem (the one about a degenerate random variable) to a probabilistic one. We next give a glorious illustration of this method, namely, use this method to prove the famous Weierstrass Approximation Theorem.

Take an arbitrary $f \in C[0,1]$. Recall that Weierstrass Approximation Theorem says that there exists a sequence $(f_m)$ of polynomials on $[0,1]$ such that $||f_m - f||_{\infty} \to 0$. In 1912, Sergei Bernstein showed that one can in fact give a formula for such a sequence:

$$f_m(t) := \sum_{k=0}^{m} \binom{m}{k} t^k (1-t)^{m-k} f \left( \frac{k}{m} \right), \quad 0 \leq t \leq 1.$$  

(Note. In approximation theory, $f_m$ is referred to as a Bernstein polynomial of degree $m$, and the fact that $||f_m - f||_{\infty} \to 0$ is called Bernstein’s Theorem.)

Fix an arbitrary real number $t$ in $[0,1]$. Let $(x_m)$ be a sequence of independent $\{0,1\}$-valued random variables on a probability space $(X, \Sigma, \mathbb{P})$ such that $\mathbb{P}\{x_m = 1\} = t$ and $\mathbb{P}\{x_m = 0\} = 1-t$ for all $m$. (By Proposition G.6.2, there is such a sequence.) Obviously, $\mathbb{E}(x_m) = t$ and $\mathbb{V}(x_m) = t(1-t)$, while an appeal to the Binomial Theorem yields $\mathbb{P}(x_1 + \cdots + x_m = k) = \binom{m}{k} t^k (1-t)^{m-k}$ for each positive integer $m$ and $k \in \{0, \ldots, m\}$. Then

$$\mathbb{E} \left( f \left( \frac{1}{m} \sum_{i \in [m]} x_i \right) \right) = f_m(t), \quad m = 1, 2, \ldots$$

To simplify the notation, let us define $y_m := \frac{1}{m}(x_1 + \cdots + x_m)$, so the expression above becomes $\mathbb{E}(f(y_m)) = f_m(t)$, for each positive integer $m$.

Let us now try to estimate $|f(t) - f_m(t)|$. For one thing,

$$|f(t) - f_m(t)| = |f(t) - \mathbb{E}(f(y_m))|$$

$$= |\mathbb{E}(f(t) - f(y_m))|$$

$$\leq \mathbb{E}(|f(t) - f(y_m)|). \quad (3)$$

Since $[0,1]$ is compact and $f$ is continuous, $f$ is uniformly continuous on $[0,1]$, and thus, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f(a) - f(b)| < \frac{\varepsilon}{2}$ for every real

\[Idea of proof.\] The Weak Law of Large Numbers implies that the probability that the random variable $y_m$ is close to $\mathbb{E}(y_m) = t$ is high for large $m$. Since $f$ is continuous, then, there is reason to expect that $f(y_m)$ is close to $f(t)$ (for large $m$) with high probability. But if so, $\mathbb{E}(f(y_m))$, which we now know to equal $f_m(t)$, should be close to $f(t)$ in the limit, exactly the sort of thing that we are after.
numbers \( a \) and \( b \) in \([0, 1]\) with \(|a - b| \leq \delta\). So, letting \( \alpha := \|f\|_\infty \), we can write
\[
\mathbb{E}(|f(t) - f(y_m)|) \leq \frac{\varepsilon}{2} p\{|t - y_m| \leq \delta\} + 2\alpha p\{|t - y_m| > \delta\} \\
\leq \frac{\varepsilon}{2} + 2\alpha p\{|t - y_m| > \delta\}.
\]
(4)

We may assume that \( \alpha > 0 \), for all is trivial when \( \alpha = 0 \). Now we invoke the Chebyshev-Bienaymé Inequality to get a handle on the number \( p\{|t - y_m| > \delta\} \). We have \( \mathbb{V}(y_m) = \frac{1}{m^2} \sum_{i \in [m]} \mathbb{V}(x_m) = \frac{1}{m} t(1 - t) \), right? Therefore,
\[
p\{|t - y_m| > \delta\} \leq \frac{\mathbb{V}(y_m)}{\delta^2} = \frac{t(1 - t)}{\delta^2 m} \leq \frac{1}{\delta^2 m}.
\]
So, if we choose an integer \( M \) large enough that \( \frac{1}{\delta^2 M} < \frac{\varepsilon}{4\alpha} \), we get \( p\{|t - y_m| > \delta\} < \frac{\varepsilon}{4\alpha} \) for every \( m \geq M \). Combining this with (4) yields
\[
\mathbb{E}(|f(t) - f(y_m)|) \leq \frac{\varepsilon}{2} + 2\alpha p\{|t - y_m| > \delta\} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\]
and, in turn, combining this with (3), we find \( |f(t) - f_m(t)| \leq \varepsilon \) for every \( m \geq M \). Since \( t \) is arbitrary and \( M \) is independent of \( t \), we thus have \( \|f_m - f\|_\infty \leq \varepsilon \) for every \( m \geq M \). Since \( \varepsilon > 0 \) is arbitrary here, the proof is complete, nice and easy!

### 2.3 Strong Law of Large Numbers

Often in applied statistical analysis one wishes to estimate the mean of a random variable. For instance, suppose we want to have a sense of the average public opinion about a particular political issue. Then we would naturally draw a random sample from the population asking each of the subjects his/her opinion. (This is just like performing the same experiment a large number of times.\(^8\)) The Weak Law of Large Numbers says simply that, for a large sample, it is likely that our sample average would approximate the true average of the population well.

To get a clearer sense of what the Weak Law of Large Numbers says (and does not say), consider again the experiment of tossing a fair coin infinitely many times, that is, take a sequence \((x_m)\) of independent \([0, 1]\)-valued random variables with

\(^8\)What would happen if I wanted to apply the Weak Law of Large Numbers at this point? Well, I would get a second-best result. By this Law, there exists a large enough positive integer \( M \) such that \( p\{|t - y_m| > \delta\} < \frac{\varepsilon}{M} \), so combining this fact with (4) and (3), we find \( |f(t) - f_m(t)| \leq \varepsilon \) for every \( m \geq M \). Why is this second-best? Because this choice of \( M \) depends on \( t \). Given that \( t \in [0, 1] \) and \( \varepsilon > 0 \) are arbitrary here, what this argument establishes is that \( f_m \to f \) pointwise. Not bad, I mind you, but what I wish to get is uniform convergence here. The way to get that is to use the Chebyshev-Bienaymé Inequality to obtain a uniform bound (with respect to \( t \)) on \( p\{|t - y_m| > \delta\} \).

\(^9\)Well, with a glitch. One would presumably not ask an individual twice (sampling without replacement), so in principle, the first experiment is not identical to (and not independent of) the second one, and so on. But if the sample space is small relative to the population size, the difficulty would not be of real substance from a practical perspective.
\[ p\{x_m = 1\} = \frac{1}{2} \] for each \( m \). The Weak Law of Large Numbers maintains that, for large (but fixed \( m \)), the relative frequency of heads is likely to be very close to \( \frac{1}{2} \). This, in turn, seems to provide a basis for interpreting the “probability” of an event as the relative frequency of that event occurring when the involved experiment is repeated a large number of times. But there is a caveat. A formal justification of this sort of an interpretation demands really something more than what the weak law is prepared to give us. In the context of our coin tossing example, for instance, what we need is that the outcome \( \omega \) of our experiment (of tossing the coin infinitely many times) is such that the relative frequency

\[
\frac{1}{m}(x_1(\omega) + \cdots + x_m(\omega))
\]

converges to \( \frac{1}{2} \) as \( m \to \infty \). Put differently, what we really need is \( \frac{1}{m}(x_1 + \cdots + x_m) \) to converge to \( \frac{1}{2} \) almost surely, but the Weak Law of Large Numbers does not yield this (and hence it is a “weak” law). As Example 1.2 shows, it is in principle possible that this relative frequency does not converge at any \( \omega \), even though it converges in probability; this would surely put in doubt the conceptual conclusion we wish to draw from the Weak Law of Large Numbers. Fortunately, this sort of a difficulty does not arise, as the following probability limit theorem, arguably the most celebrated result of modern probability theory, attests.

**The Strong Law of Large Numbers.** (Kolmogorov) For any sequence \((x_m)\) of integrable i.i.d. random variables, we have

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to_{a.s.} \mathbb{E}(x_1).
\]

In fact, a simple truncation argument shows that the integrability requirement can be relaxed to that of existence of the expectations in this result.

**Corollary 2.2.** For any sequence \((x_m)\) of i.i.d. random variables such that \( \mathbb{E}(x_1) \in \{-\infty, \infty\} \), we have

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to_{a.s.} \mathbb{E}(x_1).
\]

**Proof.** We assume, without loss of generality, \( \mathbb{E}(x_1) = \infty \). For any positive integers \( m \) and \( k \), let us put \( y_{m,k} := \min\{x_m, k\} \). Then, \( y_{m,k} = x_m \), and hence \( \mathbb{E}(y_{m,k}) = \mathbb{E}(x_m) < \infty \), and hence \( \mathbb{E}(y_{m,k}) \) is finite, for each \( m \) and \( k \). We may thus apply the Strong Law of Large Numbers to get

\[
\frac{1}{m} \sum_{i \in [m]} y_{m,k} \to_{a.s.} \mathbb{E}(y_{1,k}).
\]
As \( x_m \geq y_{m,k} \) for each \( m \) and \( k \), we have

\[
\lim \inf \frac{1}{m} \sum_{i \in [m]} x_i \geq \lim \inf \frac{1}{m} \sum_{i \in [m]} y_{m,k}
\]

and it follows that \( \lim \inf \frac{1}{m} \sum_{i \in [m]} x_i \geq_{\text{a.s.}} \mathbb{E}(y_{1,k}) \) for each \( k \). But, as \( \mathbb{E}(y_{1,k}) \) is finite and \( y_{1,k} \not\xrightarrow{\text{a.s.}} x_1 \) as \( k \to \infty \), the Monotone Convergence Theorem 2 implies \( \mathbb{E}(y_{1,k}) \to \mathbb{E}(x_1) = \infty \) as \( k \to \infty \). Thus, \( \frac{1}{m} \sum_{i \in [m]} x_i \to_{\text{a.s.}} \infty \), as we sought.

The Strong Law of Large Numbers handles the example we considered above. More generally, suppose that an experiment will be performed over and over again, and let \( S \) be an event in the experiment. If \( p(S) \) is the probability of \( S \), then the Strong Law of Large Numbers says that the relative frequency of observing \( S \) will converge to \( p(S) \) through the repetitions of the experiment. Formally, denote the probability space that corresponds to the experiment as \((X, \Sigma, p)\). Then, by the Lomnicki-Ulam Existence Theorem, the probability space that corresponds to the experiment of performing our one-stage experiment infinitely many times independently is \((X^\infty, \otimes^\infty \Sigma, p_\infty)\), where \( p_\infty \) is the product \( p \times p \times \cdots \). Define \( x_m \in \mathcal{L}^0(X^\infty, \otimes^\infty \Sigma) \) as

\[
x_m(\omega_1, \omega_2, \ldots) := \begin{cases} 1, & \text{if } \omega_m \in S \\ 0, & \text{if } \omega_m \notin S \end{cases}
\]

for each positive integer \( m \). Then \((x_m)\) is an i.i.d. sequence and we have

\[
\mathbb{E}(x_1) = \int_{X^\infty} 1_{S \times X \times X \times \cdots} d\mathbb{P}_\infty = \mathbb{P}_\infty(S \times X \times X \times \cdots) = p(S).
\]

Therefore, by the Strong Law of Large Numbers, we have

\[
\frac{1}{m} \sum_{i \in [m]} x_i \to_{\text{a.s.}} p(S).
\]

This is the formal basis of the relative frequentist interpretation of the concept of “probability.”

**Example 2.1.** Consider the game of rolling a fair dice, and suppose that your friend Jack bets \$1 on the sum of the faces coming up a prime number. If he asked you about his long-run prospects, what would be your answer?

Let us first formalize the problem by observing that here we are talking about a sequence \((x_m)\) of i.i.d. random variables with \( p\{x_1 = 1\} = \frac{15}{36} \) and \( p\{x_1 = -1\} = \frac{21}{36} \). (Note. 2, 3, 5, 7 and 11 are the only primes less than 12.) A quick computation gives \( \mathbb{E}(x_1) = \frac{1}{16} \), so you would probably tell Jack that on each game you expect him to have a negative return. But suppose Jack says “Big deal, I feel lucky today. It’s gonna be a long night!” Well, to counter this, you may attempt to compute
numbers says that, for instance, there is an
–this is the power of the laws of large numbers.
on playing this game over and over again. All this, without making any computations
1
the next section, albeit, under the additional hypothesis
of the Weak Law of Large Numbers. We shall sketch an elementary proof for it in
no need to make any computations, at least for large
low is the probability that he will end a “long” night with pro… ts. In fact, there is
probability that he will make money in the game is less than one percent. To push
the argument further, you might add, the Strong Law of Large Numbers says that the
probability that he will almost surely run out of all of his savings if he insists
on playing this game over and over again. All this, without making any computations – this is the power of the laws of large numbers. □

The proof of the Strong Law of Large Numbers is significantly harder than that
of the Weak Law of Large Numbers. We shall sketch an elementary proof for it in
the next section, albeit, under the additional hypothesis \( \mathbb{E}(x_1^4) < \infty \). The proof of
the general result will be given in the next chapter through ergodic theory. Finally,
we will offer an alternative proof in Chapter K by using the theory of martingales.

Exercise 2.4. Let \( f : \mathbb{N} \to \mathbb{R} \) be a function with \( 0 < f < \frac{1}{2} \). Consider a sequence \( (x_m) \) of \( \mathbb{N} \)-valued random variables on a probability space \((X, \Sigma, \mathbb{P})\) such that \( \mathbb{P}\{x_m = 0\} = 1 - 2f(m) \) and
\[
\mathbb{P}\left\{ x_m = -m2^m \right\} = f(m) = \mathbb{P}\{x_m = m2^m\}
\]
for each \( m \). Show that \( \mathbb{P}\left\{ \frac{1}{m} \sum_{i \in [m]} x_i \to 0 \right\} = 0 \) even though \( \mathbb{E}(x_1) = \mathbb{E}(x_2) = \cdots = 0 \).

Exercise 2.5. Take any sequence \( (x_m) \) of identically distributed integrable random variables. Assume that, for any integer \( l \geq 2 \) and any positive integers \( m_1, \ldots, m_l \) with \( m_i + 1 < m_{i+1} \) for each \( i \in [l-1] \), we have \( \prod\{x_{m_1}, \ldots, x_{m_l}\} \). Use the Strong Law of Large Numbers to prove that \( \frac{1}{m} (x_1 + \cdots + x_m) \to_{a.s.} \mathbb{E}(x_1) \).

Exercise 2.6. Let \( (x_m) \) be a sequence of i.i.d. random variables on a probability space \((X, \Sigma, \mathbb{P})\) such that the distribution of \( x_1 \) is uniform on \([0,1]\). Prove or disprove:
\[
\prod_{i \in [m]} x_i^{1/m} \to_{a.s.} x \quad \text{for some } x \in \mathcal{L}^0(X, \Sigma).
\]

Exercise 2.7. Let \( (x_m) \) be a sequence of i.i.d. random variables on a probability space \((X, \Sigma, \mathbb{P})\), and take a sequence \( (m_k) \) of \( \mathbb{N} \)-valued random variables on the same space. Use the Strong Law of Large Numbers to prove that \( m_k \to_{a.s.} \infty \) (as \( k \to \infty \)) implies
\[
\frac{1}{m_k} \sum_{i \in [m_k]} x_i \to_{a.s.} \mathbb{E}(x_1) \quad \text{as } k \to \infty.
\]

(Note. This extends the Strong Law of Large Numbers to randomly indexed sequences of i.i.d. random variables.)
2.4 Application: Borel’s Normal Number Theorem

Let \( \omega \) be a number in \([0, 1)\). Decompose \([0, 1)\) into \([0, \frac{1}{2})\) or \([\frac{1}{2}, 1)\), and notice that \( \omega \) belongs to exactly one of these intervals. If it belongs to \([0, \frac{1}{2})\), we define \( d_1(\omega) \) to be 0 and set \( d_1(\omega) := 1 \) otherwise. Now take the interval that contains \( \omega \) and decompose it into two disjoint subintervals of equal length. (For instance, if \( \omega \) belongs to \([0, \frac{1}{2})\), the decomposition yields \([0, \frac{1}{4})\) and \([\frac{1}{4}, \frac{1}{2})\).) Again, \( \omega \) must belong to exactly one of these two subintervals; we define \( d_2(\omega) \) to be 0 if \( \omega \) belongs to the subinterval that lies to the left of the other in the real line, and set \( d_2(\omega) := 1 \) otherwise. Continuing this way yields a sequence \( (d_m(\omega)) \) in \( \{0, 1\}^\infty \), which is called the dyadic expansion of \( \omega \).10

(For instance, \( (d_m(0)) = (0, 0, ...) \), \( (d_m(\frac{1}{4})) = (0, 1, 0, ...) \) and \( (d_m(\frac{7}{8})) = (1, 1, 0, ...) \).)

We say that \( \omega \) is normal if the asymptotic relative frequency of 0s (and hence 1s) is \( \frac{1}{2} \) in the dyadic expansion of \( \omega \), that is, \( \frac{1}{m} \sum i \in \{0, 1\} d_i(\omega) \to \frac{1}{2} \).

Here is a question that at first blush seems unrelated to probability theory: Is there a normal number? The answer is yes. In fact, while it is not easy to produce a concrete example of a normal number, Émile Borel showed in 1909 not only that such numbers exist, but almost all numbers in \([0, 1)\) are normal:

**Borel’s Normal Number Theorem.** \( \ell \{ \omega \in [0, 1) : \omega \text{ is normal} \} = 1 \).11

This is a very impressive result, to say the least. And yet, a minor shift in focus shows that it is but a (very) special case of the Strong Law of Large Numbers. Indeed, consider the probability space \( ([0, 1], \mathcal{B}[0, 1], \ell) \) and notice that \( d_i \) is a \( \{0, 1\} \)-valued random variable on this space for each \( i \in \mathbb{N} \).12 It is plain that \( \ell \{ d_i = 0 \} = \frac{1}{2} \) for each \( i \), that is, \( (d_m) \) are identically distributed. It is also easy to show that

\[
\ell \{ d_{i_1} = a_1, \ldots, d_{i_k} = a_k \} = 2^{-k}
\]

for every \( k \in \mathbb{N} \) and \( a_1, \ldots, a_k \in \{0, 1\} \),

that is, \( (d_m) \) is i.i.d.. Therefore, applying the Strong Law of Large Numbers yields

\[
\frac{1}{m} \sum i \in \{0, 1\} d_i \to \text{a.s.} \frac{1}{2},
\]

and this is exactly what Borel’s Normal Number Theorem says. Put differently, under its arithmetic disguise, all this theorem says is that the asymptotic relative frequency of tails (and hence of heads) in the experiment of tossing a fair coin infinitely many times is \( \frac{1}{2} \).13

---

10 By induction, we can show that \( \sum_{i \in \{0, 1\}} d_i(\omega)2^{-i} \leq \omega < \sum_{i \in \{0, 1\}} d_i(\omega)2^{-i} + 2^{-m} \), and hence \( \omega = \sum_{i=0}^{\infty} d_i(\omega)2^{-i} \), for every \( \omega \) in \([0, 1)\).

11 The fact that the set of all normal numbers is a Borel subset of \([0, 1)\) is part of the assertion.

12 So, the idea is not to concentrate on a fixed \( \omega \) and examine whether \( \omega \) is normal or not, but rather, focus on the maps \( d_i \) as functions on \([0, 1)\). As the event on which a given sequence of random variables converges is measurable – recall Section C.3 – it follows readily from this way of looking at things that the set of all normal numbers in \([0, 1)\) is Borel.

13 As an aside, let me note that it is in general quite difficult to check whether or not a given number is normal. While it is easy to verify that no rational number is normal, it is not known at present if \( \pi \) is normal or not. (The same goes for \( e \) and \( \sqrt{2} \) as well.)
Exercise 2.6. Let us say that a number $\omega$ in $[0, 1)$ is **very normal** if, for every $k \in \mathbb{N}$ and $a_1, \ldots, a_k \in \{0, 1\}$, the asymptotic relative frequency of the string $(a_1, \ldots, a_k)$ is $2^{-k}$ in the dyadic expansion of $\omega$. Show that almost every number in $[0, 1)$ is very normal.

### 2.5 Application: The Monte Carlo Method

Let $\varphi$ be an integrable real function on $[0, 1]$, and consider the problem of computing the area under the graph of $\varphi$, that is, computing

$$
\int_0^1 \varphi(t) dt.
$$

Of course, if the functional form of $\varphi$ is simple enough, we can accomplish this by using the rules of Riemann integration. If this is not the case, however, we would need to use a numerical integration technique to get an approximate answer. One way of doing this is to first choose $m$ many independent values $x_1, \ldots, x_m$ at random from $[0, 1]$ according to the uniform distribution, and then compute

$$
\frac{1}{m} \sum_{i \in [m]} \varphi(x_i)
$$

as an estimate for $\int_0^1 \varphi(t) dt$. This is the famous **Monte Carlo method** of integration (which was invented by the physicist Enrico Fermi in 1930s). But why should we believe that this method would yield reliable estimates?

Obviously, for small $m$, there is no reason to expect great accuracy from the method, but for large $m$, it should work well. After all, $\mathbb{E}(\varphi(x_1))$ equals $\int_0^1 \varphi(t) dt$, and $\varphi(x_1), \ldots, \varphi(x_m)$ are i.i.d. random variables. Therefore, the Strong Law of Large Numbers tells us that $\frac{1}{m} \sum_{i \in [m]} \varphi(x_i)$ converges to $\int_0^1 \varphi(t) dt$ almost surely, giving a sound foundation for the method of Monte Carlo integration.\(^{14}\)

In fact, we can use probability theory to say something about how large $m$ should be chosen for a reliable estimate. For instance, by the Chebyshev-Bienaymé Inequality, the probability of the event

$$
\left\{ \left| \frac{1}{m} \sum_{i \in [m]} \varphi(x_i) - \int_0^1 \varphi(t) dt \right| \geq \varepsilon \right\}
$$

is bounded above by $1/m\varepsilon^2$. Thus, to make sure that the probability that the error of our estimation is at most $\varepsilon$ is .99 (or better) we need to choose $m \geq 1/0.01\varepsilon^2$.

\(^{14}\)This method can also be used to integrate real functions of several variables. For instance, to compute $\int_0^1 \int_0^1 \varphi(s, t) ds dt$, we would sample (independently) from the uniform distribution $[0, 1]^2$. Again, the justification would be based on the Strong Law of Large Numbers.
2.6 Application: On Consistent Estimators

Suppose we are interested in the distribution of a certain characteristic in a population. Evidently, we can model this characteristic as a random variable whose distribution is given by the corresponding relative frequencies in the population. (For instance, suppose we are interested in the distribution of incomes in a given society. We can then view “income” as a random variable in the sense that, if we pick a random individual in the population, the probability that her income will be $a$ dollars is the fraction of the people in the population with income $a$.) To learn more about the nature of $x$, we would collect a random sample of size, say $m$, from the population (which, in probabilistic terms, we would interpret as running the experiment underlying our random variable $m$ many times). Our statistical inference would be based on this sample.

In statistics, this situation is modeled as follows. Let $x$ be a random variable on a probability space $(X, \Sigma, \mathbf{p})$. A random sample for $x$ is a finite collection $x_1, ..., x_m$ of i.i.d. random variables on $(X, \Sigma, \mathbf{p})$ such that $x_1 = \text{a.s.} x$. A (real-valued) statistic based on such a random sample is a random variable of the form $\varphi(x_1, ..., x_m)$, where $\varphi$ is a Borel measurable real function on $\mathbb{R} \cup \mathbb{R}^2 \cup \cdots$. (Notice that $\varphi$ can accommodate any random sample regardless of its size.) For instance, if

$$\varphi(a_1, ..., a_m) := \frac{1}{m} \sum_{i \in [m]} a_i,$$

then $\varphi(x_1, ..., x_m)$ corresponds to the statistic of the sample mean. Similarly, if

$$\varphi(a_1, ..., a_m) := \frac{1}{m} \sum_{i \in [m]} \left( a_i - \frac{1}{m} \sum_{i \in [m]} a_i \right)^2,$$

then $\varphi(x_1, ..., x_m)$ corresponds to the statistic of the sample variance.

The statistics based on a random sample are used to derive inferences about the characteristics of the random variable of interest. We would then surely wish them to satisfy certain properties. For instance, a desirable property in this regard is that of unbiasedness: We say that a statistic $\varphi(x_1, ..., x_m)$ based on the random sample $x_1, ..., x_m$ is an unbiased estimator of $\theta_x$ if

$$\mathbb{E}(\varphi(x_1, ..., x_m)) = \theta_x,$$

where $\theta_x$ is a characteristic of $x$, such as its mean or another moment. For instance, the sample mean is an unbiased estimator of $\mathbb{E}(x)$, because, for any positive integer $m$,

$$\mathbb{E} \left( \frac{1}{m} \sum_{i \in [m]} x_i \right) = \frac{1}{m} \sum_{i \in [m]} \mathbb{E}(x_i) = \mathbb{E}(x),$$

18
as $E(x_i) = E(x)$ for each $i$. By contrast, the sample variance is not an unbiased estimator of $\mathbb{V}(x)$.

The property of unbiasedness is well-defined for any random sample, regardless of its size. As such, it is said to be a small-sample property. A large-sample property of a statistic would instead be based on the limiting properties of this statistic as the sample size gets large. Of particular interest in this regard are the properties of consistency. We say that a statistic $\varphi(x_1, ..., x_m)$ based on the random sample $x_1, ..., x_m$ is a consistent estimator of a characteristic $\theta_x$ of $x$ if

$$\text{p-} \lim \varphi(x_1, ..., x_m) = \theta_x,$$

and that it is a strongly consistent estimator of $\theta_x$ if

$$\text{p}\{\varphi(x_1, ..., x_m) \to \theta_x\} = 1.$$

Laws of large numbers are indispensable tools for determining the consistency properties of a given statistic. We consider two illustrations of this next.

**Example 2.2.** The sample mean is a strongly consistent estimator of $E(x)$, provided that $E(x)$ is finite. This is the same thing as saying that

$$\frac{1}{m} \sum_{i \in [m]} x_i \xrightarrow{\text{a.s.}} E(x),$$

when $E(x)$ is finite. And as such, it is none other than a restatement of the Strong Law of Large Numbers.

**Example 2.3.** The sample variance is a strongly consistent estimator of $\mathbb{V}(x)$, provided that both $E(x)$ and $E(x^2)$ are finite. This is the same thing as saying that

$$\frac{1}{m} \sum_{i \in [m]} \left( x_i - \frac{1}{m} \sum_{i \in [m]} x_i \right)^2 \xrightarrow{\text{a.s.}} \mathbb{V}(x),$$

when $\mathbb{V}(x)$ is finite. To see this, note first that our claim is equivalent to

$$\frac{1}{m} \sum_{i \in [m]} x_i^2 - \left( \frac{1}{m} \sum_{i \in [m]} x_i \right)^2 \xrightarrow{\text{a.s.}} \mathbb{V}(x),$$

as the left-hand sides of the two expressions above are easily verified to be one and the same. But as $x_1^2, x_2^2, ...$ are i.i.d. and $E(x^2)$ is finite, the Strong Law of Large Numbers

---

15The bias of this estimator is, however, negligible when the sample size $m$ is large.
Numbers entails that $\frac{1}{m} \sum_{i \in [m]} x_i^2 \to_{a.s.} \mathbb{E}(x^2)$. Similarly, $\frac{1}{m} \sum_{i \in [m]} x_i \to_{a.s.} \mathbb{E}(x)$ and hence $(\frac{1}{m} \sum_{i \in [m]} x_i^2) \to_{a.s.} \mathbb{E}(x)^2$. It follows that

$$\frac{1}{m} \sum_{i \in [m]} x_i^2 - \left( \frac{1}{m} \sum_{i \in [m]} x_i \right)^2 \to_{a.s.} \mathbb{E}(x^2) - \mathbb{E}(x)^2 = \mathbb{V}(x),$$

as we sought. \hfill \Box

Exercise 2.8. Consider the real map $\varphi$ defined on $\mathbb{R} \cup \mathbb{R}^2 \cup \cdots$ by

$$\varphi(a_1, \ldots, a_m) := \frac{1}{m - 1} \sum_{i \in [m]} \left( a_i - \frac{1}{m} \sum_{i \in [m]} a_i \right)^2.$$

Show that $\varphi(x_1, \ldots, x_m)$ is an unbiased and strongly consistent estimator of $\mathbb{V}(x)$, provided that both $\mathbb{E}(x)$ and $\mathbb{E}(x^2)$ are finite.

### 2.7 Application: On Convergence of Empirical Distributions

Consider the previous setting in which we used random samples to derive inferences about a random variable of interest, say $x \in \mathcal{L}^0(X, \Sigma)$. Suppose this time that we wish to use our random samples to estimate the entire distribution of $x$. The idea is to view the values of a random sample $x_1, \ldots, x_m$ for $x$ as a realization of these random variables at a particular outcome $\omega$ in $X$. Then, the probability distribution that puts mass $1/m$ at each $x_i(\omega)$ – this is called an empirical distribution for $x$ – seems like a reasonable estimator for the distribution of $x$. In particular, we would expect this distribution to approximate that of $x$ fairly well for large $m$. But observe that this approximation is parametric over $\omega$. (Two different random samples of size $m$ corresponds to two different realizations of $x_1, \ldots, x_m$, thereby yielding two different empirical distributions.) The question is if we can be sure that empirical distributions for a random variable would approximate the distribution of that random variable well for all $\omega$. As we shall show presently, the Strong Law of Large Numbers yields a very nice answer to this question.

Let us investigate the problem in abstract terms. Let $Y$ be a compact metric space, and $x$ a $Y$-valued random variable on a probability space $(X, \Sigma, \mathbb{P})$. Let $x_1, x_2, \ldots$ be i.i.d. $Y$-valued random variables on $(X, \Sigma, \mathbb{P})$ with $x_1 =_{a.s.} x$. For any positive integer $m$ and outcome $\omega \in X$, we define the simple probability measure $\mathbb{P}_{m, \omega} \in \Delta(Y)$ by $\mathbb{P}_{m, \omega}(\{x_i(\omega)\}) = \frac{1}{m}$ for each $i \in [m]$. The measure $\mathbb{P}_{m, \omega}$ is called the empirical distribution for $x$ based on the random sample $x_1, \ldots, x_m$ at $\omega$.

Notice that, for any $\omega \in X$ and $\varphi \in C(Y)$, we have

$$\int_X \varphi \mathbb{P}_{m, \omega} = \frac{1}{m} \sum_{i \in [m]} \varphi(x_i(\omega)).$$

But $\varphi \circ x_1, \varphi \circ x_2, \ldots$ are i.i.d. random variables on $(X, \Sigma, \mathbb{P})$, so, by the Strong Law of Large Numbers,

$$\frac{1}{m} \sum_{i \in [m]} \varphi \circ x_i \to_{a.s.} \int_X \varphi \circ x \mathbb{P}.$$

That is, there is a set $S(\varphi) \in \Sigma$ such that $\mathbb{P}(S(\varphi)) = 0$ and

$$\int_X \varphi \mathbb{P}_{m, \omega} \to \int_X \varphi \circ x \mathbb{P} \quad \text{for every } \omega \in X \setminus S(\varphi).$$
Now, since \( Y \) is compact, \( C(Y) \) is separable, and hence there is a countable dense set \( \{ \varphi_1, \varphi_2, \ldots \} \) in \( C(Y) \). Letting \( S := S(\varphi_1) \cup S(\varphi_2) \cup \cdots \), we find \( p(S) = 0 \) and

\[
\int_X \varphi_i d p_{m, \omega} \to \int_X \varphi_i \circ x d p \quad \text{for every } i \in \mathbb{N} \text{ and } \omega \in X \setminus S.
\]

Since \( \{ \varphi_1, \varphi_2, \ldots \} \) is dense in \( C(Y) \), this means that

\[
\int_X \varphi d p_{m, \omega} \to \int_X \varphi \circ x d p \quad \text{for every } \varphi \in C(Y) \text{ and } \omega \in X \setminus S,
\]

that is, \( p\{ \omega \in X : p_{m, \omega} \xrightarrow{w} p_x \} = 1 \). In fact, contrary to how it looks, compactness of \( Y \) is not essential here. With a bit of help from real analysis, we can relax this property to separability.

**Varadarajan’s Theorem.** Let \( Y \) be a separable metric space, and \( x, x_1, x_2, \ldots \) i.i.d. \( Y \)-valued random variables on a probability space \( (X, \Sigma, p) \). Then,

\[
p\left\{ \omega \in X : p_{m, \omega} \xrightarrow{w} p_x \right\} = 1.
\]

**Exercise 2.10.** Prove Varadarajan’s Theorem by using Exercise E.4.8.

In the case of (real-valued) random variables, we can establish something significantly stronger. Indeed, if \( x, x_1, x_2, \ldots \) are i.i.d. random variables on a probability space \( (X, \Sigma, p) \), \( \omega \in X \), and \( F_{m, \omega} \) is the distribution function induced by \( p_{m, \omega} \), then Varadarajan’s Theorem and Proposition E.1.7 entail that the probability of the event (i.e., the set of all \( \omega \in X \) such that \( F_{m, \omega}(t) \to F_x(t) \)) for every \( t \) at which \( F \) is continuous is one. Furthermore, if \( t \) is a discontinuity point of \( F \), then applying the Strong Law of Large Numbers to the sequence \( 1_{\left(-\infty, t\right]} \circ x_1, 1_{\left(-\infty, t\right]} \circ x_2, \ldots \), we find that \( p\{ \omega : F_{m, \omega}(t) \to F_x(t) \} = 1 \). Conclusion: \( p\{ \omega : F_{m, \omega} \to F_x \} = 1 \). And this is not the end of the story. We can in fact show that \( F_{m, \omega} \to F_x \) uniformly with probability one.

**Exercise 2.11.** (Glivenko-Cantelli Theorem) If \( x, x_1, x_2, \ldots \) are i.i.d. random variables on a probability space \( (X, \Sigma, p) \), then \( p\{ \omega : F_{m, \omega} \to F_x \text{ uniformly} \} = 1 \).

(a) Prove this result in the case where \( x \) is \((0, 1)\)-valued and \( p_x = \ell \).

(b) Prove the general result by using the observation noted in Remark B.5.4.

## 3 The Borel-Cantelli Lemmas

### 3.1 The First Lemma

A problem that arises frequently in probability limit theory is the calculation of the probability of the upper limit of a certain sequence of independent events. This task is often simplified by the very important fact that such an event occurs either with probability 0 or with probability 1! We shall prove this curious result in this subsection, and point to some of its applications.

**Convergence Part of the Borel-Cantelli Lemma**
We divide the statement of the said 0-1 law into two parts. Remarkably, the first part – sometimes called the *convergence part* of the Borel-Cantelli Lemma – does not even require the independence hypothesis.\(^{16}\)

**The Borel-Cantelli Lemma 1.** Let \((X, \Sigma, p)\) be a probability space, and \((S_m)\) a sequence of events in \(\Sigma\) such that \(\sum_{i=1}^{\infty} p(S_i) < \infty\). Then, \(p(\lim\sup S_m) = 0\).

**Proof.** Since \(\lim\sup S_m \subseteq S_k \cup S_{k+1} \cup \cdots\) for every positive integer \(k\), Boole’s Inequality implies

\[
p(\lim\sup S_m) \leq p(S_k \cup S_{k+1} \cup \cdots) \leq \sum_{i=k}^{\infty} p(S_i).
\]

But, since \(\sum_{i=1}^{\infty} p(S_i)\) converges, we have \(p(S_k) + p(S_{k+1}) + \cdots \to 0\) as \(k \to \infty\) (Exercise A.3.9). The claim is thus proved upon letting \(k \to \infty\). \(\blacksquare\)

“Using” the Borel-Cantelli Lemma 1

Since the almost sure convergence of a sequence of random variables can be established by checking whether or not the probabilities of the upper limits of certain sequences of events vanish (Lemma 1.2), the Borel-Cantelli Lemma 1 often proves useful when computing the almost sure limit of a sequence of random variables. Here is an illustration.

**Example 3.1.** Let \((x_m)\) be a sequence of identically distributed random variables on a probability space \((X, \Sigma, p)\). We wish to prove the following:

\[
\mathbb{E}(|x_1|) < \infty \quad \text{implies} \quad \frac{1}{m} |x_m| \to_{a.s.} 0.
\]

By Lemma 1.2, our problem is to show that

\[
p \left( \lim\sup \left\{ \frac{1}{m} |x_m| > \varepsilon \right\} \right) = 0 \quad \text{for every} \quad \varepsilon > 0.
\]

This would follow from the Borel-Cantelli Lemma 1 if we could show that

\[
\sum_{i=1}^{\infty} p \left\{ \frac{1}{\varepsilon} |x_i| > i \right\} < \infty \quad \text{for every} \quad \varepsilon > 0.
\]

But, as \(x_i\)'s are identically distributed, we have \(\mathbb{E}(\frac{1}{\varepsilon} |x_1|) \geq \sum_{i=1}^{\infty} p \left\{ \frac{1}{\varepsilon} |x_i| > i \right\}\) by Corollary D.3.2. Thus, (5) holds whenever \(\mathbb{E}(|x_1|)\) is finite, and we are done. \(\square\)

---

\(^{16}\)The Borel-Cantelli Lemmas were stated for independent random variables by Emile Borel in 1909, but Borel’s proof contained some flaws. Francesco Cantelli in 1917 gave a correct proof for the result, and noted that one direction of the lemma does not require the variables be independent.
Proposition 3.1. Let \( (X, \Sigma, \mathbf{p}) \) be a probability space, and \((S_m)\) a sequence in \(\Sigma\). Prove: If \(\sum_{i=1}^{\infty} \mathbf{p}(S \cap S_i) < \infty\), then \(\mathbf{p}(\lim \sup S_m) \leq 1 - \mathbf{p}(S)\).

Exercise 3.1. Let \((X, \Sigma, \mathbf{p})\) be a probability space, \((S_m)\) a sequence in \(\Sigma\). Prove: There exists a sequence \((a_m)\) of positive integers such that \(\frac{1}{a_m} x_m \rightarrow_{a.s.} 0\).

Exercise 3.2. Let \((X_m)\) be a sequence of random variables on a probability space \((X, \Sigma, \mathbf{p})\). Prove: There exists a sequence \((a_m)\) of positive integers such that \(\frac{1}{a_m} x_m \rightarrow_{a.s.} 0\).

Exercise 3.3. Let \((X_m)\) be a sequence of random variables on a probability space \((X, \Sigma, \mathbf{p})\), and \((a_m)\) a real sequence with \(\sum_{i=1}^{\infty} a_i < \infty\). Prove: If \(\sum_{i=1}^{\infty} \mathbf{p}\{x_i \geq a_i\} < \infty\), then \(\sum_{i=1}^{\infty} x_i\) converges almost surely.

Exercise 3.4. Let \((x_m)\) and \((y_m)\) be two sequences of random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that \(\sum_{i=1}^{\infty} \mathbf{p}\{x_i \neq y_i\} < \infty\). (Note. \((x_m)\) and \((y_m)\) are said to be equivalent in the sense of Khinchine.) Show that \(\sum_{i=1}^{\infty} x_i\) converges almost surely if and only if \(\sum_{i=1}^{\infty} y_i\) converges almost surely.

Exercise 3.5. Let \((a_m)\) and \((b_m)\) be two sequences of nonnegative real numbers such that \(\sum_{i=1}^{\infty} a_i < \infty\) and \(\sum_{i=1}^{\infty} b_i < \infty\). Let \((x_m)\) be a sequence of random variables on a probability space \((X, \Sigma, \mathbf{p})\). Prove: If \(\mathbf{p}\{|x_{m+1} - x_m| > b_m\} < a_m\) for each \(m\), then \((x_m)\) converges almost surely.

As another application, we show how the Borel-Cantelli Lemma 1 may be used to establish a special case of the Strong Law of Large Numbers.

Exercise 3.6. (Borel’s Strong Law of Large Numbers) Let \((x_m)\) be a sequence of i.i.d. random variables on a probability space \((X, \Sigma, \mathbf{p})\). Assume that \(\mathbb{E}(x_1) = 0\) and \(\mathbb{E}(x_1^2) < \infty\).

(a) Use Proposition G.2.1 to establish the following:

\[
\mathbb{E}\left(\left(\sum_{i\in[m]} x_i\right)^4\right) = \mathbb{E}\left(\sum_{i,j,k,l\in[m]} x_i x_j x_k x_l\right) = m \mathbb{E}(x_1^4) + 3(m^2 - m)(\mathbb{E}(x_1^2))^2.
\]

(b) Use the Chebyshev-Bienaymé Inequality and the Borel-Cantelli Lemma 1 to show that

\[
\mathbf{p}\left(\lim \sup \left\{\sum_{i\in[m]} x_i \left| m \epsilon \right\}\right\} > m \epsilon\right) = 0 \quad \text{for every } \epsilon > 0.
\]

(c) Conclude that \(\frac{1}{m}(x_1 + \cdots + x_m) \rightarrow_{a.s.} 0\).

Convergence in Probability vs. Almost Sure Convergence, Again

We have seen earlier that a sequence of random variables that converges in probability need not converge almost surely (Example 1.2). Remarkably, however, such a sequence is sure to possess a subsequence that converges almost surely. As we shall see later, this is a very useful observation that often facilitates deriving certain types of almost sure convergence theorems. We now prove this result as an application of the Borel-Cantelli Lemma 1.

Proposition 3.1. Let \(Y\) be a separable metric space, and \(x, x_1, x_2, \ldots\) \(Y\)-valued random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that \(\mathbf{p}\)-lim \(x_m = x\). Then, there exists a strictly increasing sequence \((m_k)\) of positive integers such that \(x_{m_k} \rightarrow_{a.s.} x\).
**Proof.** Take any strictly decreasing real sequence \((\varepsilon_m)\) in \((0,1)\) with \(\sum_{i=1}^{\infty} \varepsilon_i < \infty\). Define
\[
m_1 := \min\{m \in \mathbb{N} : \mathbf{p}\{d_Y(x_i, x) > \varepsilon_1\} \leq \varepsilon_1 \text{ for all } i \geq m\}
\]
and
\[
m_{k+1} := \min\{m \in \{m_k + 1, \ldots\} : \mathbf{p}\{d_Y(x_i, x) > \varepsilon_{k+1}\} \leq \varepsilon_{k+1} \text{ for all } i \geq m\}
\]
for every positive integer \(k\). Since \(\mathbf{p}\)-lim \(x_m = x\) by hypothesis, each of these numbers is well-defined, and of course, \((m_k)\) is a strictly increasing sequence in \(\mathbb{N}\). Furthermore, by construction, \(\mathbf{p}\{d_Y(x_{m_k}, x) > \varepsilon_k\} \leq \varepsilon_k\) for each \(k \geq 1\), so
\[
\sum_{k=1}^{\infty} \mathbf{p}\{d_Y(x_{m_k}, x) > \varepsilon_k\} \leq \sum_{k=1}^{\infty} \varepsilon_k < \infty.
\]
Thus, by the Borel-Cantelli Lemma 1, we have
\[
\mathbf{p}(\limsup\{d_Y(x_{m_k}, x) > \varepsilon_k\}) = 0.
\]
Since \(\varepsilon_k \searrow 0\) (because \(\sum_{k=1}^{\infty} \varepsilon_k\) is finite), it follows from this observation that
\[
\mathbf{p}(\limsup\{d_Y(x_{m_k}, x) > \varepsilon\}) = 0 \quad \text{for every } \varepsilon > 0.
\]
By Lemma 1.2, then, we are done.

**Exercise 3.7.** Let \((x_m)\) be a sequence of random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that \(x_1 \leq x_2 \leq \cdots\). Show that if \(\mathbf{p}\)-lim \(x_m = x\) for some \(x \in \mathcal{L}^0(X, \Sigma)\), then \(x_m \rightarrow_{a.s.} x\).

**Exercise 3.8.** Let \(Y\) be a separable metric space, and \(x, x_1, x_2, \ldots\) \(Y\)-valued random variables on a probability space \((X, \Sigma, \mathbf{p})\). Show that \(\mathbf{p}\)-lim \(x_m = x\) if and only if every subsequence of \((x_m)\) has a subsequence that converges to \(x\) almost surely.

**Exercise 3.9.** Let \((x_m)\) be a sequence of random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that \(|x_m| \leq_{a.s.} y\) for each \(m\). Show that if \(\mathbf{p}\)-lim \(x_m = x\) for some \(x \in \mathcal{L}^0(X, \Sigma)\), then \(\mathbb{E}(x_m) \rightarrow \mathbb{E}(x)\).

**A Generalization of the Borel-Cantelli Lemma 1**

We conclude this subsection with a famous generalization of the Borel-Cantelli Lemma 1, which was established in 1961 by Ole Barndorff-Nielsen.

**The Barndorff-Nielsen Lemma.** Let \((X, \Sigma, \mathbf{p})\) be a probability space, and \((S_m)\) a sequence of events in \(\Sigma\) such that
\[
\mathbf{p}(S_m) \rightarrow 0 \quad \text{and} \quad \sum_{i=1}^{\infty} \mathbf{p}(S_i \cap (X \setminus S_{i+1})) < \infty.
\]
Then, \( p(\limsup S_m) = 0 \).

**Proof.** An outcome \( \omega \) belongs to \( \limsup S_m \) if either it belongs to all but finitely many of the sets \( S_1, S_2, \ldots \) or (exclusive) it belongs to infinitely many of the sets \( S_1, S_2, \ldots \) as well as infinitely many of the sets \( X \setminus S_1, X \setminus S_2, \ldots \). But the latter occurs if \( \omega \in S_m \cap (X \setminus S_{m+1}) \) for infinitely many \( m \). Thus,

\[
\limsup S_m = \liminf S_m \cup (\limsup (S_m \cap (X \setminus S_{m+1}))).
\]

Therefore, by Lemma 1.1,

\[
p(\limsup S_m) = p(\liminf S_m) + p(\limsup (S_m \cap (X \setminus S_{m+1})))
\leq \liminf p(S_m) + p(\limsup (S_m \cap (X \setminus S_{m+1}))),
\]

and the proof is completed applying the present hypotheses and the Borel-Cantelli Lemma 1. \( \blacksquare \)

### 3.2 The Second Lemma

We now concentrate on the converse of the Borel-Cantelli Lemma 1. It is easily seen that we need an additional hypothesis in this regard. For instance, the sequence of events \(([0, \frac{1}{m})])\) in the Borel probability space \(([0, 1], \mathcal{B}[0, 1], \ell)\) satisfies

\[
\ell(0, 1) + \ell(0, \frac{1}{2}) + \cdots = 1 + \frac{1}{2} + \cdots = \infty
\]

whereas \( \ell(\limsup [0, \frac{1}{m})) = \ell\{0\} = 0 \). Moreover, in general, there is no reason for the probability of observing the upper limit of a sequence of events to be 0 or 1. For instance, take any probability space \((X, \Sigma, p)\) and an event \( S \in \Sigma \). Then, the limsup of the event sequence \((S, S, \ldots)\) is \( S \), and hence, the probability of observing the terms of this sequence infinitely often, that is, \( p(\limsup S) \), equals \( p(S) \), which, obviously, may be a number in \((0, 1)\).

**Divergence Part of the Borel-Cantelli Lemma**

What goes wrong in the examples above is that the events that they work with are not independent. It is a truly remarkable fact that independence would dispense with such examples right away. That is, in the case of independent events, the converse of the Borel-Cantelli Lemma 1 is true.

**The Borel-Cantelli Lemma 2.** Let \((X, \Sigma, p)\) be a probability space and \((S_m)\) be a sequence of independent events in \( \Sigma \). If \( \sum_{i=1}^{\infty} p(S_i) = \infty \), then \( p(\limsup S_m) = 1 \).

**Proof.** Note first that \( \{X \setminus S_1, X \setminus S_2, \ldots\} \) is an independent sequence (Exercise G.1.2). Consequently, for any positive integers \( k \) and \( K \) such that \( K \geq k + 1 \),

\[
p\left(\bigcap_{i=k}^{\infty} X \setminus S_i\right) \leq p\left(\bigcap_{i=k}^{K} X \setminus S_i\right) = \prod_{i=k}^{K} \left(1 - p(S_i)\right) \leq e^{-p(S_k) + \cdots + p(S_K)}
\]

25
where the final step follows from the inequality \(1 - a \leq e^{-a}\) which is valid for any real number \(a\) between 0 and 1.\(^{17}\) Then letting \(K \to \infty\), we find \(p(\cap_{i \geq k} X \setminus S_i) = 0\) for each \(k \geq 1\), because \(\sum_{i=1}^{\infty} p(S_i) = \infty\). Thus, by Boole’s Inequality,

\[
p(\lim \inf X \setminus S_m) = p\left(\bigcap_{k=1}^{\infty} \bigcap_{i=k}^{\infty} X \setminus S_i\right) \leq \sum_{k=1}^{\infty} p\left(\bigcap_{i=k}^{\infty} X \setminus S_i\right) = 0.
\]

As \(\lim \inf X \setminus S_m = X \setminus \lim \sup S_m\), then, we have \(p(\lim \sup S_m) = 1\).

**The Borel 0-1 Law**

The Borel-Cantelli Lemmas 1 and 2 jointly provide a complete picture about the likelihood of the occurrence of the upper limit of a sequence of independent events in a given probability space. If \((S_m)\) is such a sequence, then we have

\[
p(\lim \sup S_m) := \begin{cases} 1, & \text{if } p(S_1) + p(S_2) + \cdots = \infty \\ 0, & \text{otherwise}. \end{cases}
\]

This fact – some authors refer to it as the **Borel 0-1 Law** – has numerous applications within probability theory.

**“Using” the Borel-Cantelli Lemma 2**

**Example 3.2.** Let \(Y\) be a metric space, and \((x_m)\) a sequence of independent \(Y\)-valued random variables on a probability space \((X, \Sigma, p)\). For any \(x \in \mathcal{L}^0(X, \Sigma)\), Lemma 1.2 tells us that \(x_m \to_a.s. x\) iff

\[
p\left(\lim \sup \{d_Y(x_m, x) > \varepsilon\}\right) = 0 \quad \text{for every } \varepsilon > 0.
\]

By the Borel 0-1 Law, we obtain an alternative characterization: \(x_m \to_a.s. x\) iff

\[
\sum_{i=1}^{\infty} p\{d_Y(x_i, x) > \varepsilon\} < \infty \quad \text{for every } \varepsilon > 0.
\]

Given the nature of the particular problem one is interested in, this characterization may be easier to check than either the previous one or using the definition of almost sure convergence directly. \(\square\)

**Example 3.3.** Let \((x_m)\) be a sequence of independent random variables on a probability space \((X, \Sigma, p)\), such that

\[
\sum_{i=1}^{\infty} p\{|x_i| > i\} = \infty.
\]

\(^{17}\)The map \(a \mapsto e^{-a} + a - 1\) is increasing on \([0, 1]\) and takes value 0 at 0.
Let us show that, where \( y_k := x_1 + \cdots + x_k \) for every positive integer \( k \), the sequence \( (\frac{1}{m}y_m) \) does not converge to 0 almost surely. Thanks to the Borel-Cantelli Lemma 2, this is easy. After all, this lemma implies \( p\{|x_m| > m \text{ infinitely often}\} = 1 \) here. But for any integer \( m \geq 2 \), we have \( |x_m| \leq |y_m| + |y_{m-1}| \) by the Triangle Inequality, and hence \( \{|x_m| > m \text{ infinitely often}\} \) is contained within the event
\[
\{|y_m| > m \text{ infinitely often}\} \cup \{|y_{m-1}| > m \text{ infinitely often}\}.
\]
But the two events here are one and the same – think about it! – so,
\[
1 = p\{|x_m| > m \text{ infinitely often}\} \leq p\{\frac{1}{m}|y_m| > 1 \text{ infinitely often}\}.
\]
Thus, not only that \( (\frac{1}{m}y_m) \) does not converge to 0 almost surely, the probability of the event \( \{\frac{1}{m}y_m \to 0\} \) is zero! \( \square \)

**Example 3.4.** Consider the following question:

What is the probability that two consecutive heads will come up infinitely often in the repeated tossing of a fair coin?

To answer this question, we adopt the model introduced in Example B.3.2 and denote by \( S_m \) the event that heads come up in the \( m \)th and \( (m + 1) \)th trial, that is,
\[
S_m := \{(\omega_1, \omega_2, \ldots) \in \{0, 1\}^\infty : \omega_m = \omega_{m+1} = 1\}.
\]
Notice that \( S_m \) and \( S_{m+1} \) are not independent events for any \( m \), but \( (S_{2m}) \) is a sequence of independent events. (Why?) Moreover, \( p(S_2) + p(S_4) + \cdots = \infty \), and hence, by the Borel-Cantelli Lemma 2, \( p(\limsup S_m) \geq p(\limsup S_{2m}) = 1 \). The answer to our question is thus \( 1. \)\( ^{18} \)

**Example 3.5.** (More on Record Values) Consider the setup we introduced in Section G.2.7. Given a sequence \( (x_m) \) of continuous i.i.d. random variables on a probability space \((X, \Sigma, p)\), let us pose the following two questions:

1. What is the probability that we shall observe a record infinitely many times along the sample path of \( (x_m) \)?

2. What is the probability that we shall observe two consecutive records infinitely many times along the sample path of \( (x_m) \)?

(Any guesses?) The Borel 0-1 Law allows us to answer these questions with ease. Take question (1) first. In terms of the notation of Section G.2.7, we are interested in computing \( p\{R_m \text{ infinitely often}\} \). But we know from our discussion in Section

\(^{18}\)Not impressed? Fine, then tell me what is the probability of observing one million heads in a row infinitely often. The same argument shows that – thanks, again, to the Borel-Cantelli Lemma 2 – this is also 1!
Lemma 1.2.7 that \( R_1, R_2, \ldots \) are independent events such that \( p(R_m) = \frac{1}{m} \) for each positive integer \( m \). Therefore, \( p(R_1) + p(R_2) + \cdots = \infty \), and hence, by the Borel-Cantelli Lemma 2, the probability of observing infinitely many records along the sample paths of \((x_m)\) is one.

Let us now take on question (2). Consider the following events:

\[ S_k := \{ \omega \in X : \text{both } x_k(\omega) \text{ and } x_{k+1}(\omega) \text{ are record values} \} \]

where \( k \) is any positive integer.\(^\text{19}\) We wish to compute \( p\{ S_m \text{ infinitely often} \} \). Observe that

\[ p(S_m) = p(R_m \cap R_{m+1}) = p(R_m)p(R_{m+1}) = \frac{1}{m(m+1)} \]

for each \( m \geq 1 \). Consequently,

\[ \sum_{i=1}^{\infty} p(S_i) = \sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \sum_{i=1}^{\infty} \left( \frac{1}{i} - \frac{1}{i+1} \right) = \lim_{m \to \infty} \left( 1 - \frac{1}{m} \right) < \infty. \]

Therefore, by the Borel-Cantelli Lemma 1, we conclude that it is with probability zero that we shall observe two consecutive records infinitely many times along the sample path of \((x_m)\). \( \Box \)

**Exercise 3.10.** Let \((x_m)\) be a sequence of i.i.d random variables on a probability space \((X, \Sigma, p)\). Show that \( E(|x_1|) = \infty \) implies \( p\left( \frac{1}{m} |x_m| \to \infty \right) = 1 \).

**Exercise 3.11.** Let \((X, \Sigma, p)\) be a probability space, and \((S_m)\) a sequence of independent events in \( \Sigma \) such that \( p(S_m) < 1 \) for each \( m \).

(a) Prove that \( p(\limsup S_m) = 1 \) iff \( p(S_1 \cup S_2 \cup \cdots) = 1 \).

(b) Using the probability space \(([0,1], B[0,1], \ell)\) and the event sequence \(((\frac{1}{2}, 1), [0, \frac{1}{2}), [0, \frac{1}{2}), \ldots)\), show that the “if” part of the previous claim is false without the independence hypothesis.

**Exercise 3.12.** (Bauer) Let \( x_2, x_3, \ldots \) be independent random variables on a probability space \((X, \Sigma, p)\) such that

\[ p\{x_m = -m\} = \frac{1}{2m \ln m} = p\{x_m = m\} \quad \text{and} \quad p\{x_m = 0\} = 1 - \frac{1}{m \ln m}, \]

for each positive integer \( m \geq 2 \). Use Example 3.3 to conclude that \( \frac{1}{m} \sum_{i \in [m]} x_i \) does not converge to 0 almost surely. Next, use Markov’s Weak Law of Large Numbers to show that \( \frac{1}{m} \sum_{i \in [m]} x_i \) converges to 0 in probability. (Thus the sequence \((x_2, x_3, \ldots)\) satisfies the conclusion of the Weak, but not the Strong, Law of Large Numbers.)

\(^{19}\) These events are not independent, but that’s okay, for I’m going to work with the Borel-Cantelli Lemma 1 here.
We conclude with two exercises that illustrate how one may be able to weaken the independence hypothesis in the Borel-Cantelli Lemma 2.\textsuperscript{20}

**Exercise 3.13.** Let \((X, \Sigma, p)\) be a probability space, and \((S_m)\) a sequence in \(\Sigma\) with \(p(S_i \cap S_j) \leq p(S_i)p(S_j)\) for every distinct positive integers \(i\) and \(j\). (The events \(S_i\) are thus negatively correlated.) Show that if \(\sum \infty p(S_i) = \infty\), then we have \(p(\lim \sup S_m) = 1\).

**Exercise 3.14.** (The Erdös-Rényi Theorem) Let \((X, \Sigma, p)\) be a probability space, and \((S_m)\) a sequence of pairwise independent events in \(\Sigma\). Show that if \(\sum \infty p(S_i) = \infty\), then we have \(p(\lim \sup S_m) = 1\).

### 4 Convergence of Series of Random Variables

#### 4.1 Maximal Inequalities of Kolmogorov and Ottaviani

Most convergence theorems for sums of infinitely many random variables are built on some form of a probability inequality that gives an upper bound for the probability of the event that the partial sums of the given random variables are large. The following inequality, which was obtained by Andrei Kolmogorov in 1928, and which generalizes the Chebyshev-Bienaymé Inequality, is a prime example of such probability inequalities.

**Kolmogorov’s Maximal Inequality.** Given a positive integer \(m\), let \(x_1, \ldots, x_m\) be independent random variables on a probability space \((X, \Sigma, p)\) such that \(E(x_1) = \cdots = E(x_m) = 0\). Then,

\[
p\left\{ \left| \sum_{i \in [k]} x_i \right| \geq a \text{ for some } k \in [m] \right\} \leq \frac{1}{a^2} \sum_{i \in [m]} V(x_i)
\]

for any real number \(a > 0\).\textsuperscript{21}

The method of proof we shall use for Kolmogorov’s Maximal Inequality is a standard technique of probability theory. Succinctly put, the idea is to consider summing randomly many of our random variables in a suitable manner to be able to decompose the events about the maximal partial sums into disjoint events the probabilities of which are easier to compute. (Another way of looking at this problem will be outlined in Chapter L.)

**Proof of Kolmogorov’s Maximal Inequality.** Let us assume that \(m \geq 2\) (for otherwise the result reduces to the Chebyshev-Bienaymé Inequality), and define \(y_k := x_1 + \cdots + x_k\) for each \(k \in [m]\). Throughout the argument \(a\) is taken to be an arbitrarily fixed positive real number.

We define the map \(M : X \to [m + 1] \) as

\[
M(\omega) := \begin{cases} 
\min\{k : |y_k| \geq a\}, & \text{if } |y_k| \geq a \text{ for some } k \in [m] \\
m + 1, & \text{otherwise.}
\end{cases}
\]

\textsuperscript{20}There are many other generalizations of the Borel-Cantelli Lemmas. See Kochen and Stone (1964) and Petrov (2002), for instance.

\textsuperscript{21}The left-hand side of this inequality can be written as \(p\{ \max\{|x_1 + \cdots + x_k| : k \in [n]\} \geq a\}\). This is the reason why one refers to the said inequality as a “maximal” inequality.
Clearly, $M$ is a simple random variable on $(X, \Sigma, \mathbf{p})$. The key observation is that
\[
\{|y_k| \geq a \text{ for some } k \in [m]\} = \{y_M^2 \geq a^2\}
\]
(6)
as is easily checked.\footnote{Note that $y_M$ is a randomly indexed random variable. Indeed, we have
\[y_M(\omega) = x_1(\omega) + \cdots + x_M(\omega), \quad \omega \in X,\]
that is, $y_M$ corresponds to the sum of a random number of the random variables $x_1, \ldots, x_m$.}
The advantage of this formulation is that Markov’s Inequality applies to $y_M$:
\[
P\{y_M^2 \geq a^2\} \leq \frac{\mathbb{E}(y_M^2)}{a^2}.
\]
(7)
Besides, given that $x_1, \ldots, x_m$ are independent and $\mathbb{E}(x_1) = \cdots = \mathbb{E}(x_m) = 0$, we have
\[
\sum_{i \in [m]} \mathbb{V}(x_i) = \mathbb{V} \left( \sum_{i \in [m]} x_i \right) = \mathbb{E}(y_m^2).
\]
By (6) and (7), therefore, we will be done if we can show that
\[
\mathbb{E}(y_M^2) \leq \mathbb{E}(y_m^2).
\]
To this end, define the event $S_i := \{M = i\}$ for each $i \in [m]$, and note that
\[
\mathbb{E}(y_M^2) = \mathbb{E}(y_1^21_{S_1}) + \cdots + \mathbb{E}(y_m^21_{S_m}).
\]
Thus, to prove $\mathbb{E}(y_M^2) \leq \mathbb{E}(y_m^2)$, it is enough to show that
\[
\int_{S_i} y_i^2d\mathbf{p} \leq \int_{S_i} (y_i + (x_{i+1} + \cdots + x_m))^2d\mathbf{p}
\]
for each $i \in [m-1]$, which would, in turn, follow from
\[
\int_{S_i} y_i(x_{i+1} + \cdots + x_m)d\mathbf{p} \geq 0
\]
for each $i \in [m-1]$. Well, this one is easy to see. After all, the independence of $x_1, \ldots, x_m$ implies that of $y_i1_{S_i}$ and $x_{i+1} + \cdots + x_m$ — right? — and hence,
\[
\int_{S_i} y_i(x_{i+1} + \cdots + x_m)d\mathbf{p} = \int_{S_i} y_id\mathbf{p} \int_X (x_{i+1} + \cdots + x_m)d\mathbf{p} = \left( \int_{S_i} y_id\mathbf{p} \right) \left( \mathbb{E}(x_{i+1}) + \cdots + \mathbb{E}(x_m) \right) = 0
\]
for each $i \in [m-1]$. The theorem is now proved. \hfill \blacksquare

Kolmogorov’s Maximal Inequality provides an upper bound for a tail probability of the maximum of partial sums of finitely many random variables by using the variance of the total sum of these random variables. The following inequality, which was established by Giorgio Ottaviani in 1939, is similar to this, but it uses as the upper bound also a tail probability for the sum of all the random variables at hand. The proof of this result exploits again the “random sum” technique we used above.
Ottaviani’s Maximal Inequality. Given a positive integer \( m \geq 2 \), let \( x_1, \ldots, x_m \) be independent random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that

\[
P \left\{ \left| \sum_{i=j+1}^{m} x_i \right| > a \right\} \leq \beta, \quad j \in [m-1]
\]  

(8)

for some real numbers \( a > 0 \) and \( \beta \in (0,1) \). Then,

\[
P \left\{ \sum_{i \in [k]} x_i \geq 2a \text{ for some } k \in [m] \right\} \leq \frac{1}{1-\beta} P \left\{ \sum_{i \in [m]} x_i > a \right\}.
\]

Proof. Let \( y_k := x_1 + \cdots + x_k \) for each \( k \in [m] \), and define \( M : X \to [m+1] \) as

\[M(\omega) := \begin{cases} \min\{k : |y_k| \geq 2a\}, & \text{if } |y_k| \geq 2a \text{ for some } k \in [m] \\ m+1, & \text{otherwise.} \end{cases}\]

The key observation here is that

\[\{ |y_m| > a \} \supseteq \{ M = k \text{ and } |y_m - y_k| \leq a \}\]

for each \( k \in [m] \). Consequently, as \( \{ M = k \} \) and \( \{ |x_{k+1} + \cdots + x_m| \leq a \} \) are independent events – because the former belongs to \( \sigma(x_1, \ldots, x_k) \) and the latter to \( \sigma(x_{k+1}, \ldots, x_m) \) – we have

\[
P \{ |y_m| > a \} \geq \sum_{k \in [m]} P \{ M = k \} P \{ |y_m - y_k| \leq a \} \geq (1 - \beta) \sum_{k \in [m]} P \{ M = k \},
\]

where the second inequality follows from (8). Since \( \sum_{k \in [m]} P \{ M = k \} \text{ equals } P \{ |y_k| \geq 2a \text{ for some } k \in [m] \} \) by definition of \( M \), we are done. \( \blacksquare \)

In the next section, we shall use Ottaviani’s Maximal Inequality to establish a fundamental result about the almost sure convergence of an infinite series of independent random variables.

Exercise 4.1. (Etemadi’s Inequality) Given a positive integer \( m \), let \( x_1, \ldots, x_m \) be independent random variables on a probability space \((X, \Sigma, \mathbf{p})\). Prove that, for any \( a > 0 \),

\[
P \{ \max \{ |y_k| : k \in [m] \} > 3a \} \leq 3 \max \{ P \{ |y_k| > 3a \} : k \in [m] \},
\]

where \( y_k := x_1 + \cdots + x_k \) for each \( k \in [m] \).

Exercise 4.2. (The Hájek-Rényi Inequality) Given a positive integer \( m \), let \( x_1, \ldots, x_m \) be independent random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that \( \mathbb{E}(x_1) = \cdots = \mathbb{E}(x_m) = 0 \). Then,

\[
P \left\{ \alpha_k \left| \sum_{i \in [k]} x_i \right| \geq a \text{ for some } k \in [m] \right\} \leq \frac{1}{a^2} \sum_{i \in [m]} \alpha_i^2 \mathbb{V}(x_i)
\]

for any real numbers \( \alpha_1 \geq \cdots \geq \alpha_m > 0 \) and \( a > 0 \).
4.2 Lévy’s Theorem

We have seen in Section 1.2 that almost sure convergence is more demanding than convergence in probability, which is the cause of the wedge between the weak and strong laws of large numbers. Remarkably, however, this difference dissipates in the context of infinite series of independent random variables. That is, such a series is almost surely convergent if it converges in probability.

**Lévy’s Theorem.** Let \((x_m)\) be a sequence of independent random variables on a probability space \((X, \Sigma, \mathbf{p})\). Then, \(\sum_{i=1}^{\infty} x_i\) converges almost surely if, and only if, it converges in probability.

As we shall see, Ottaviani’s Inequality makes it quite easy to prove this result. All we need is the following auxiliary fact.

**The Cauchy Criterion for Almost Sure Convergence.** Let \(Y\) be a Polish space, and \((x_m)\) a sequence of \(Y\)-valued random variables on a probability space \((X, \Sigma, \mathbf{p})\) such that, for every \(\varepsilon > 0\),

\[
P\{d_Y(x_m, x_k) \geq \varepsilon \text{ for some } k > m\} \to 0 \quad \text{as } m \to \infty.
\]

Then, \(x_m \to_a.s. x\) for some \(Y\)-valued random variable \(x\) on \((X, \Sigma, \mathbf{p})\).

**Proof.** For any positive integers \(m\) and \(n\), define

\[
A_{mn} := \{d_Y(x_m, x_k) < \frac{1}{n} \text{ for every } k > m\} \quad \text{and} \quad A_n := A_{1n} \cup A_{2n} \cup \cdots.
\]

(Thanks to Example B.5.4, we have \(A_{mn} \in \Sigma\) for each \(m\) and \(n\).) As \(A_{1n} \subseteq A_{2n} \subseteq \cdots\), we have \(P(A_{mn}) \searrow P(A_n)\), for every positive integer \(n\). As our hypothesis implies that \(P(A_{mn}) \to 1\), we may thus conclude that \(P(A_n) = 1\) for every \(n = 1, 2, \ldots\). Then, \(P(A) = 1\), where \(A := A_1 \cap A_2 \cap \cdots\). But \(\omega \in A\) means that \((x_m(\omega))\) is a Cauchy sequence in \(Y\). Since \(Y\) is complete, therefore, we may define the map \(x : X \to Y\) as

\[
x(\omega) := \begin{cases} 
\lim x_m(\omega), & \text{if } \omega \in A \\
\nu, & \text{otherwise,}
\end{cases}
\]

where \(\nu\) is an arbitrarily fixed point in \(Y\). It remains to check that \(x\) is a \(Y\)-valued random variable \(x\) on \((X, \Sigma, \mathbf{p})\). We leave this step as an exercise.

We are now fully prepared to prove Lévy’s Theorem.

**Proof of Lévy’s Theorem.** Given Proposition 1.2, we need only to prove the “if” part of the assertion. Take any real numbers \(\varepsilon > 0\) and \(\delta \in (0, 1)\). We wish to find a positive integer \(M\) with

\[
P\left\{\sum_{i=m+1}^{k} x_i \geq \varepsilon \text{ for some } k > m\right\} < \delta
\]

for every \(m \geq M\). (In view of the Cauchy Criterion for Almost Sure Convergence, this is enough to complete our proof.)

**Claim.** There exists a positive integer \(M\) such that

\[
P\left\{\sum_{i=s}^{t} x_i \geq \frac{\varepsilon}{2}\right\} \leq \frac{\delta}{2} \quad \text{for every } t \geq s \geq M.
\]
Proof. Let \( y_k := x_1 + \cdots + x_k \) for every positive integer \( k \). We are given that \( p\text{-lim } y_m = y \) for some random variable on \((X, \Sigma, p)\). But, for any positive integers \( s \) and \( t \),

\[
\left\{ |y_t - y_s| > \frac{\varepsilon}{2} \right\} \subseteq \left\{ |y_t - y| \geq \frac{\varepsilon}{4} \right\} \cup \left\{ |y - y_s| \geq \frac{\varepsilon}{4} \right\}
\]

whence

\[
p \left\{ |y_t - y_s| > \frac{\varepsilon}{2} \right\} \leq p \left\{ |y_t - y| \geq \frac{\varepsilon}{4} \right\} + p \left\{ |y - y_s| \geq \frac{\varepsilon}{4} \right\}.
\]

As \( p\text{-lim } y_m = y \), there is an \( M \in \mathbb{N} \) such that \( p \left\{ |y_m - y| \geq \varepsilon/4 \right\} \leq \delta/4 \) for every integer \( m \geq M \), and hence follows (9). \( \| \)

Now, let \( M \) be as found in our claim above. Then, for any integers \( m \) and \( K \) with \( K > m \geq M \),

\[
p \left\{ \left| \sum_{i=j+m}^{K} x_i \right| > \frac{\varepsilon}{2} \right\} \leq \frac{\delta}{2} \quad \text{for every } j = 0, \ldots, K - m.
\]

Then, by Ottaviani’s Inequality (applied to the random variables \( x_m, \ldots, x_K \)) and (9),

\[
p \left\{ \sum_{i=m}^{k} x_i \geq \varepsilon \text{ for some } k = m, \ldots, K \right\} \leq \frac{1}{1 - \delta/2} p \left\{ \left| \sum_{i=m}^{K} x_i \right| > \frac{\varepsilon}{2} \right\}
\]

\[
< \frac{\delta/2}{1 - \delta/2}
\]

\[
< \delta,
\]

and our proof is complete. \( \qed \)

Lévy’s Theorem is quite impressive already, but we can in fact do even better than this. First, notice that our random variables in this result need not be real-valued. Indeed, the same argument (replacing the absolute value sign with the norm sign where appropriate) tells us that, for any Banach space \( Y \) and independent \( Y \)-valued random variables \( x_1, x_2, \ldots \) on a probability space \((X, \Sigma, p)\), there exists a \( Y \)-valued random variable \( y \) on \((X, \Sigma, p)\) such that

\[
p \left\{ \left\| \sum_{i \in [m]} x_i - y \right\| \to 0 \right\} = 1
\]

iff \( \sum_{i} x_i \) converges in probability. What is more, one can replace the term “in probability” in this statement with the term “in distribution.” That is, almost sure convergence, convergence in probability and convergence in distribution are equivalent concepts in the case of infinite series of random variables on a given probability space.\(^{23}\)

Exercise 4.2.\(^{41}\) Show that the independence hypothesis can be omitted in the statement of Lévy’s Theorem if we have \( x_i \geq 0 \) for each \( i = 1, 2, \ldots \)

\(^{23}\)The proof of this strengthening of Lévy’s Theorem is beyond the scope of this text. If you are familiar with characteristic functions, have a look at Ito and Nisio (1968).
4.3 The Kolmogorov Convergence Criterion

We now wish to apply Lévy’s Theorem to obtain an easy-to-check sufficient condition for the almost sure convergence of an infinite series of independent random variables. The following is one of the most brilliant results of asymptotic probability theory.

The Kolmogorov Convergence Criterion. Let \((x_m)\) be a sequence of independent random variables such that \(\mathbb{E}(x_1) = \mathbb{E}(x_2) = \cdots = 0\) and \(\sum_{i=1}^{\infty} \mathbb{E}(x_i^2) < \infty\). Then,

\[
\sum_{i=1}^{\infty} x_i \text{ converges almost surely.}
\]

**Proof.** Let \(y_m := x_1 + \cdots + x_m\) for every positive integer \(m\). Observe that, for every positive integers \(k\) and \(l\) with \(l > k\), we have

\[
\mathbb{E}\left((y_l - y_k)^2\right) = \mathbb{V}(y_l - y_k) = \sum_{i=k+1}^{l} \mathbb{V}(x_i) = \sum_{i=k+1}^{l} \mathbb{E}(x_i^2),
\]

where the first equality holds because \(\mathbb{E}(y_l - y_k) = 0\) (as \(\mathbb{E}(x_i) = 0\) for each \(i\)), the second because of independence, and the third because \(\mathbb{E}(x_i) = 0\) for each \(i\). Given that \(\sum_{i=1}^{\infty} \mathbb{E}(x_i^2) < \infty\), letting \(k \to \infty\) here, therefore, we find that \((y_m)\) is a Cauchy sequence in \(L^2(X, \Sigma, p)\). Since \(L^2(X, \Sigma, p)\) is a complete metric space – recall the Riesz-Fischer Theorem – then,

\[
\int_X (y_m - y)^2 d\mathbb{p} \to 0 \quad \text{for some} \; y \in L^2(X, \Sigma, p).
\]

But then, for any \(\varepsilon > 0\), the Chebyshev-Bienaymé Inequality implies

\[
\mathbb{P}\{|y_m - y| > \varepsilon\} \leq \frac{1}{\varepsilon^2} \int_X (y_m - y)^2 d\mathbb{p} \to 0,
\]

so we conclude that \(\mathbb{p}\text{-lim} y_m = y\). Applying Lévy’s Theorem completes the proof.

**Corollary 4.1.** Let \((x_m)\) be a sequence of independent random variables such that \(\sum_{i=1}^{\infty} \mathbb{V}(x_i) < \infty\). Then,

\[
\sum_{i=1}^{\infty} (x_i - \mathbb{E}(x_i)) \text{ converges almost surely.}
\]

**Proof.** Let \(z_i := x_i - \mathbb{E}(x_i)\) for every positive integer \(i\). Then we have \(\bigcap\{z_1, z_2, \ldots\}\) and \(\mathbb{E}(z_1) = \mathbb{E}(z_2) = \cdots = 0\). Moreover, \(\sum_{i=1}^{\infty} \mathbb{V}(z_i^2) = \sum_{i=1}^{\infty} \mathbb{V}(x_i) < \infty\), so our assertion follows from the Kolmogorov Convergence Criterion.

The following set of exercises provides several illustrations of how one would use the Kolmogorov Convergence Criterion in practice.

In the following set of exercises \((x_m)\) stands for a sequence of independent random variables on a given probability space \((X, \Sigma, p)\).

**Exercise 4.3.** Assume that \(x_1, x_2, \ldots\) are identically distributed and \(\mathbb{P}\{x_1 = -1\} = 1/2 = \mathbb{P}\{x_1 = 1\}\). Does \(\sum_{i=1}^{\infty} x_i / i\) converge almost surely?
Exercise 4.4. (*Cantor Distribution*) Assume that \(x_1, x_2, \ldots\) are identically distributed and 
\[ p\{x_1 = 0\} = 1/2 = p\{x_1 = 2\}. \] Does \(\sum_{i=1}^{\infty} x_i/3^i\) converge almost surely?

Exercise 4.5. Let \((\lambda_n)\) be a real sequence such that \(\inf\{\lambda_1, \lambda_2, \ldots\} > 0\), and assume that \(x_i\) is exponentially distributed with parameter \(\lambda_i > 0\), \(i = 1, 2, \ldots\) Does \(\sum_{i=1}^{\infty} x_i/i^2\) converge almost surely?

Exercise 4.6. We say that a random variable \(x\) is **symmetrically distributed** if \(x = a.s. -x\). Assume that \(x_1, x_2, \ldots\) are identically and symmetrically distributed. Prove that \(\sum_{i=1}^{\infty} x_i/i\) converges almost surely if \(x_1\) is integrable.

Exercise 4.7. Assume that \(x_1, x_2, \ldots\) are symmetrically distributed, and we have

\[ \sup \left\{ \mathbb{E} \left( \left( \sum_{i \in [m]} x_i \right)^2 \right) : m = 1, 2, \ldots \right\} < \infty. \]

Show that \(\sum_{i=1}^{\infty} x_i\) converges almost surely.

Exercise 4.8. Assume that \(\mathbb{E}(x_i^2) < \infty\) for each \(i = 1, 2, \ldots\), and that there exists a random variable \(x\) on \((X, \Sigma, \mathbb{P})\) with \(\mathbb{E}(\sum_{i \in [m]} (x_i - x)^2) \to 0. \) Show that \(\sum_{i=1}^{\infty} x_i\) converges almost surely.

Exercise 4.9. (The Three-Series Theorem) Assume that there is a real number \(c > 0\) such that each of the following series converge:

\[ \sum_{i=1}^{\infty} p\{|x_i| > c\}, \quad \sum_{i=1}^{\infty} \mathbb{E}(x_i 1_{\{|x_i| \leq c\}}), \quad \sum_{i=1}^{\infty} \mathbb{V}(x_i 1_{\{|x_i| \leq c\}}). \]

Show that \(\sum_{i=1}^{\infty} x_i\) converges almost surely.\(^{24}\)

Exercise 4.10. Assume that \(x_1, x_2, \ldots\) are identically distributed, \(\mathbb{E}(x_1) = 0, \mathbb{E}(x_1^2) = 1,\) and for some real number \(c > 0, p\{|x_1| > c\} = 0.\) Let \((a_m)\) be a sequence of positive real numbers with \(\sum_{i=1}^{\infty} a_i x_i \leq \infty. \) Show that \(\sum_{i=1}^{\infty} x_i\) converges almost surely.

Exercise 4.11. Assume that \(x_1, x_2, \ldots\) are identically distributed and \(p\{x_1 = -1\} = \alpha\) and \(p\{x_1 = 1\} = 1 - \alpha,\) for some real number \(\alpha\) in \((0, 1).\) Let \((a_m)\) be a sequence of nonnegative real numbers. What exactly must \((a_m)\) satisfy so that \(\sum_{i=1}^{\infty} a_i x_i\) converges almost surely?

Exercise 4.12. Prove the Kolmogorov Convergence Criterion by using the Kolmogorov Maximal Inequality instead of Lévy’s Theorem.

5 0-1 Laws

5.1 Kolmogorov’s 0-1 Law

The Strong Law of Large Numbers says that, for any sequence \((x_m)\) of integrable i.i.d. random variables, the sequence \(\left( \frac{1}{m} (x_1 + \cdots + x_m) \right)\) of partial averages of these random variables converges to \(\mathbb{E}(x_1)\) almost surely. It turns out that part of this conclusion

\(^{24}\)The converse of this result is also true, that is, the convergence of these three series is necessary and sufficient for almost sure convergence of \(\sum_{i=1}^{\infty} x_i.\) (This is also due to Kolmogorov by the way. Who else?) The proof of the necessity part of this statement is a bit involved, however.

35
would remain valid if we dropped the hypotheses that $x_m$'s have identical distributions, and that they are integrable. Curiously, only on the basis of the independence of $x_1, x_2, \ldots$ we can be certain that both $(x_m)$ and $(\frac{1}{m}(x_1 + \cdots + x_m))$ will either converge almost surely to a constant random variable, or diverge almost surely. This fact is extremely useful in studying the long run behavior of a sequence of independent random variables (although which of the two alternatives is actually true is usually quite difficult to discern).

**Tail Events**

The following concept plays a key role in the analysis of sequences of independent random variables.

**Definition.** Let $Y$ be a metric space, $(X, \Sigma)$ a measurable space, and $(x_m)$ a sequence of $Y$-valued random variables on $(X, \Sigma)$. The tail $\sigma$-algebra of $(x_m)$ is defined as

$$T(x_m) := \bigcap_{i=1}^{\infty} \sigma(x_i, x_{i+1}, \ldots).$$

Any member of $T(x_m)$ is called a **tail event associated with** $(x_m)$.

Given any sequence of $Y$-valued random variables, recall that $\sigma(x_m, x_{m+1}, \ldots)$ is the collection of all events the occurrence or non-occurrence of which can be determined on the basis of observing $x_m, x_{m+1}, \ldots$ alone. Consequently, the occurrence or non-occurrence of an event in $T(x_m)$ can be determined without any reference to the random variables $x_1, \ldots, x_m$. As $m$ is arbitrary here, therefore, we may think of a tail event associated with $(x_m)$ as one that does not rely on any finite subset of $\{x_1, x_2, \ldots\}$. (Replacing finitely many of the $x_i$'s with some other random variables (on the same measurable space), for instance, would not alter $T(x_m)$.25) This intuition suggests that tail events have a lot to do with the asymptotic behavior of $(x_m)$. The following examples show that this is indeed the case.

**Example 5.1.** Let $(X, \Sigma)$ be a measurable space, and $x_1, x_2, \ldots \in L^0(X, \Sigma)$. Consider the event that the sum of the terms of $(x_m)$ converges to a finite number, that is, $S := \{\sum_{i}^{\infty} x_i \text{ converges}\}$. We wish to show that $S$ is a tail event associated with $(x_m)$.

For any integers $k$ and $m$ with $k \geq m$, define

$$y_m := \limsup_{i=m}^{k} \sum_{i=m}^{k} x_i \quad \text{and} \quad z_m := \liminf_{i=m}^{k} \sum_{i=m}^{k} x_i.$$

25 Quiz. Is $\{x_m \to x_1\}$ a tail event associated with $(x_m)$?
Now fix, arbitrarily, a positive integer \( m \), and set \( T(m) := \sigma(x_m, x_{m+1}, \ldots) \). As \( x_m, \ldots, x_k \) are \( T(m) \)-measurable, \( x_m + \cdots + x_k \in \mathcal{L}^0(X, T(m)) \) for every \( k \) with \( k \geq m \). It follows that both \( y_m \) and \( z_m \) are \( \mathbb{R} \)-valued random variables on \((X, T(m))\). Therefore, both \( A_m := \{ y_m < \infty \} \) and \( B_m := \{ z_m > -\infty \} \), and hence \( A_m \cap B_m \), belong to \( T(m) \). This also implies that

\[
 w_m := (y_m - z_m)1_{A_m \cap B_m}
\]

is a random variable on \((X, T(m))\) — why? — and hence \( C_m := \{ w_m = 0 \} \in T(m) \).

The key observation here is that we have \( A_1 = A_m, B_1 = B_m \) and \( C_1 = C_m \). (Why?) Thus: \( A_1 \cap B_1 \cap C_1 \in T(m) \). As \( m \) is arbitrarily chosen in \( \mathbb{N} \) and \( S = A_1 \cap B_1 \cap C_1 \), we may then conclude that \( S \in T(x_m) \), as we sought.

Exercise 5.1. Let \((X, \Sigma)\) be a measurable space, and \((x_m)\) a sequence in \( \mathcal{L}^0(X, \Sigma) \). Prove that \( \{ \lim x_m > 0 \} \) and \( \{ \frac{1}{m} \sum_{i\in[m]} x_i \to 1 \} \) are tail events associated with \((x_m)\).

Exercise 5.2. Let \( Y \) be a metric space, \((X, \Sigma)\) a measurable space, and \((x_m)\) a sequence of \( Y \)-valued random variables on \((X, \Sigma)\). Prove that, for any \((S_m) \in \sigma(x_1) \times \sigma(x_2) \times \cdots \), \( \lim \sup S_m \) and \( \lim \inf S_m \) are tail events associated with \((x_m)\).

The Kolmogorov 0-1 Law

An amazing result of probability theory says that any tail event associated with a sequence of independent random variables is either almost sure to occur or almost sure not to occur. We will use this result later in proving the Strong Law of Large Numbers.

**Kolmogorov’s 0-1 Law.** Let \( Y \) be a metric space and \((x_m)\) a sequence of independent \( Y \)-valued random variables on a probability space \((X, \Sigma, p)\). If \( S \) is a tail event associated with \((x_m)\), then \( p(S) \in \{0, 1\} \).

**Proof.** By the Grouping Lemma (of Section G.1.2), \( \sigma(x_1), \ldots, \sigma(x_{m-1}) \) and \( \sigma(x_m, x_{m+1}, \ldots) \) are independent for every integer \( m \geq 2 \).\(^{26}\) Since \( T(x_m) \subseteq \sigma(x_m, x_{m+1}, \ldots) \), it follows that \( \sigma(x_1), \ldots, \sigma(x_{m-1}) \) and \( T(x_m) \) are independent for every integer \( m \geq 2 \). This, in turn, implies that \( \sigma(x_1), \sigma(x_2), \ldots \) and \( T(x_m) \) are independent. (Yes?) By the Grouping Lemma, then, \( \sigma(\bigcup \{ \sigma(x_i) : i \in \mathbb{N} \}) \) — that is, \( \sigma(x_1, x_2, \ldots) \) — and \( T(x_m) \) are independent. Since, by definition, \( T(x_m) \subseteq \sigma(x_1, x_2, \ldots) \), it follows that \( T(x_m) \) is independent of itself. So, if \( S \in T(x_m) \), then \( p(S) = p(S \cap S) = p(S)^2 \), and hence \( p(S) \in \{0, 1\} \). \( \blacksquare \)

So, if \((x_m)\) is a sequence of independent random variables, the probability that \( \sum_{i=1}^{\infty} x_i \) converges is either zero or one. Similarly, \( \frac{1}{m} \sum_{i\in[m]} x_i \) is either almost surely convergent or almost surely divergent. (Compare with the Strong Law of Large Numbers.)

\(^{26}\)I’m using here the fact that \( \sigma(x_m, x_{m+1}, \ldots) = \sigma(\sigma(x_m) \cup \sigma(x_{m+1}) \cup \cdots) \).
Corollary 5.1. If $(x_m)$ is a sequence of independent random variables on a probability space $(X, \Sigma, \mu)$, then either there exists an extended real number $a$ such that $x_m \to a$ almost surely or $(x_m)$ diverges almost surely. The same holds also for the sequence $(\frac{1}{m} \sum_{i \in [m]} x_i)$.

Exercise 5.3. Prove Corollary 5.1.

Exercise 5.4. Let $(x_m)$ be a sequence of i.i.d. random variables on a probability space $(X, \Sigma, \mu)$. Prove or disprove: $\mathbb{P}\{\sum_{i}^\infty x_i \text{ converges} \} = 1$ if $x_1 = a$.

Exercise 5.5. Let $(x_m)$ be a sequence of independent random variables on a probability space $(X, \Sigma, \mu)$ such that $\mathbb{P}\{x_m = 0\} = 1 - 2^{-m}$ and $\mathbb{P}\{x_m = 1\} = 2^{-m}$ for each $m$. Show that $0 < \mathbb{P}\{x_m = 1 \text{ for some } m\} < 1$. What is going on?

Exercise 5.6. Let $(x_m)$ be a sequence of independent random variables on a probability space $(X, \Sigma, \mu)$, and $y := \varphi(x_1, x_2, ...)$ for some $\varphi : \mathbb{R}^\infty \to \mathbb{R}$. Prove: If $y$ is $T(x_m)$-measurable – in this case we say that $y$ is a tail function associated with $(x_m)$ – then the distribution function $F_y$ of $y$ satisfies:

$$F_y(t) = \begin{cases} 0, & \text{if } t < \inf\{s : \mathbb{P}_y\{y \leq s\} = 1\} \\ 1, & \text{otherwise.} \end{cases}$$

That is, $y$ almost surely constant.

Exercise 5.7. Prove: If $(x_m)$ is a sequence of independent random variables on a probability space, then there exist extended real numbers $a$ and $b$ such that $\liminf x_m = a$ and $\limsup x_m = b$.

5.2 The Hewitt-Savage 0-1 Law

Exchangeable Events

Let $Y$ be a separable metric space, and $(x_m)$ a sequence of $Y$-valued random variables on a probability space. By Exercise G.6.3, $\mathcal{B}(Y^\infty) = \bigotimes^\infty \mathcal{B}(Y)$. Therefore, by Exercise G.2.3, for every event $S$ in $\sigma(x_1, x_2, ...)$, there is an event $S^\bullet$ in $\bigotimes^\infty \mathcal{B}(Y)$ such that $S = \{(x_1, x_2, ...) \in S^\bullet\}$.

We say that an event $S$ in $\sigma(x_1, x_2, ...)$ is an exchangeable event associated with $(x_m)$ if there is an $S^\bullet$ in $\bigotimes^\infty \mathcal{B}(Y)$ such that

$$\{(x_1, x_2, ...) \in S^\bullet\} = S = \{(x_{\rho(1)}, x_{\rho(2)}, ...) \in S^\bullet\}$$

for every finite permutation $\rho$ on $\mathbb{N}$. In words, this means that the occurrence or non-occurrence of $S$ is not affected when finitely many of the random variables in the sequence $(x_m)$ change places. The collection of all exchangeable events associated with $(x_m)$ is a sub-$\sigma$-algebra of $\sigma(x_1, x_2, ...)$, as can easily be checked. We denote this $\sigma$-algebra as $\mathcal{E}(x_m)$.

---

27Put more precisely, $\sigma(x_1, x_2, ...)$ consists exactly the sets of the form $\{(x_1, x_2, ...) \in A\}$ where $A \in \mathcal{B}(Y) \otimes \mathcal{B}(Y) \otimes \cdots$, a result which obtains by combining the said two exercises.

28Reminder. A finite permutation $\rho$ on $\mathbb{N}$ is a bijective self-map on $\mathbb{N}$ such that $\rho(i) = i$ for all but finitely many $i \in \mathbb{N}$. 

38
Exercise 5.8. A real function $f$ on $\mathbb{R}^k$ is said to be symmetric if $f(a_1, \ldots, a_n) = f(a_{\rho(1)}, \ldots, a_{\rho(k)})$ for every real numbers $a_i$, $i \in [k]$. For any sequence $(x_m)$ of random variables and $k \in \mathbb{N}$, let $\mathcal{E}(k)$ be the $\sigma$-algebra generated by all maps of the form $f(x_1, \ldots, x_k)$ where $f$ is a symmetric Borel measurable real map on $\mathbb{R}^k$. Show that $\mathcal{E}(x_m) = \mathcal{E}(1) \cap \mathcal{E}(2) \cap \cdots$.

The following examples illustrate why exchangeable events associated with a random sequence is of interest.

Example 5.2. Let $Y$ be a separable metric space, and $(x_m)$ a sequence of $Y$-valued random variables on a probability space. We wish to show that a tail event associated with $(x_m)$ is exchangeable, that is, $\mathcal{T}(x_m) \subseteq \mathcal{E}(x_m)$. (Intuitively, this is because a tail event associated with $(x_m)$ does not depend on any finitely many of initial terms of the sequence $(x_m)$, so it is, trivially, exchangeable with respect to these random variables.)

Let $S$ be a tail event associated with $(x_m)$. For any positive integer $m$, because $S$ belongs to $\sigma(x_{m+1}, x_{m+2}, \ldots)$, there exists a $T^*_{m+1} \in \otimes_\infty \mathcal{B}(Y)$ such that $S = \{(x_{m+1}, x_{m+2}, \ldots) \in T^*_{m+1}\}$. We then define $S^* := \limsup(Y^m \times T^*_{m+1})$, and observe that $S = \{x_m \in S^*\}$. It is easy to check that $S^*$ is symmetric under finite permutations, that is, a sequence $(\nu_m)$ belongs to $S^*$ iff so does $(\nu_{\rho(m)})$, for every finite permutation $\rho$ on $\mathbb{N}$. It follows that $S$ is an exchangeable event associated with $(x_m)$.  

Insight. Every tail event associated with a random sequence is exchangeable.

Example 5.3. Let $(x_m)$ be a sequence of random variables on a probability space. Then, $\{x_1 = x_2 = \cdots\}$ is an exchangeable, but in general not a tail, event associated with $(x_m)$. Similarly, the events $\{\sum_1^{\infty} x_i = 1\}$ and $\{\limsup \sum_1^{m} x_i \leq 3\}$ are exchangeable, but in general not tail, events associated with $(x_m)$.  

The Hewitt-Savage 0-1 Law

Kolmogorov’s 0-1 Law says that a tail event associated with a sequence of independent random variables has probability 0 or 1. This need not be true for an exchangeable event associated with this sequence. However, it turns out that the conclusion remains intact for exchangeable events as well, provided that the involved random variables are i.i.d. This is the content of the following result which was proved by Edwin Hewitt and Leonid Savage in 1955.

The Hewitt-Savage 0-1 Law. Let $Y$ be a separable metric space and $(x_m)$ a sequence of i.i.d. $Y$-valued random variables on a probability space $(X, \Sigma, p)$. If $S$ is an exchangeable event associated with $(x_m)$, then $p(S) \in \{0, 1\}$.

\footnote{We could as well use $\limsup(Y^m \times T^*_{m+1})$ for $S^*$ here.}
Proof. (Throughout this proof we denote \((x_m)\) by \(x\), and \((x_{\rho(m)})\) by \(x_{\rho}\) for any permutation \(\rho\) on \(\mathbb{N}\).) Take any \(S \in \mathcal{E}_x\), and pick any \(S^*\) in \(\otimes^\infty \mathcal{B}(Y)\) such that (10) holds for every finite permutation \(\rho\) on \(\mathbb{N}\). We now recall Exercise B.2.6, and invoke the definition of \(\otimes^\infty \mathcal{B}(Y)\) and the Approximation Lemma (of Section B.2.2) to conclude that there is a sequence \((T_n)\) of cylinder sets in \(Y^\infty\) such that \(d_{p_x}(T_n, S^*) \to 0\), from which it follows that \(p\{x \in T_n\} \to p\{x \in S^*\}\). Furthermore, the same is true for the sequence \(\{x^\rho \in T_n\}\) of events, that is, \(p\{x^\rho \in T_n\} \to p\{x \in S^*\}\), for any finite permutation \(\rho\) on \(\mathbb{N}\), because, by

\[
p\{x^\rho \in T_n\} \triangle \{x \in S^*\} \quad = \quad p\{x^\rho \in T_n\} \triangle \{x \in S^*\} \\
\quad = \quad p_{x^\rho}(T_n \triangle S^*) \\
\quad = \quad p_x(T_n \triangle S^*) \\
\quad \to 0,
\]

where we used the exchangeability of \(S\) to get the first equality, and the i.i.d.

hypothesis (which entails \(p_x = p_{x^\rho}\) as we have seen in Proposition G.6.3) to get the third.

Now, for each \(n\), because \(T_n\) is a cylinder set, there is a positive integer \(k_n\) and \(A_n \in Y^{k_n}\) such that \(T = A_n \times Y \times Y \times \cdots\). We consider the finite permutation \(\rho_n\) on \(\mathbb{N}\) that map any \(i \in [k_n]\) to \(k_n + i\), and any \(i \in [2k_n]\setminus[k_n]\) to \(i - k_n\), leaving any \(i \in \mathbb{N}\setminus[2k_n]\) fixed. Besides, for \(\Lambda_n := \{x \in T_n\} \cap \{x^\rho \in T_n\}\), we have

\[
p(\Lambda_n) = p\{(x_1, \ldots, x_{k_n}) \in A_n \text{ and } (x_{k_n+1}, \ldots, x_{2k_n}) \in A_n\} \\
\quad = p\{(x_1, \ldots, x_{k_n}) \in A_n\} p\{(x_{k_n+1}, \ldots, x_{2k_n}) \in A_n\} \\
\quad = p\{x \in T_n\} p\{x^\rho \in T_n\}
\]

by independence, for each \(n = 1, 2, \ldots\). Thus: \(p(\Lambda_n) \to p(S)^2\). On the other hand, by

the triangle inequality for the semimetric \(d_p\),

\[
d_p(\Lambda_n, S) \quad \leq \quad p(\Lambda_n \triangle \{x \in T_n\}) + d_p(\{x \in T_n\}, S) \\
\quad \leq \quad p(\{x \in T_n\} \triangle \{x^\rho \in T_n\}) + d_p(\{x \in T_n\}, S) \\
\quad \leq \quad d_p(\{x \in T_n\}, S) + d_p(S, \{x^\rho \in T_n\}) + d_p(\{x \in T_n\}, S)
\]

for each \(n\), so letting \(n \to \infty\), we find \(d_p(\Lambda_n, S) \to 0\), which entails \(p(\Lambda_n) \to p(S)\). It follows that \(p(S)^2 = p(S)\), and we are done. \(\blacksquare\)

The biggest shortcoming of the Hewitt-Savage 0-1 Law is its hypothesis that \(p_{x_1} = p_{x_2} = \cdots\). This hypothesis cannot be omitted in the statement of the result. (Example. For a sequence \((x_m)\) of independent random variables such that \(p_{x_1}(1) = \frac{1}{2} = p_{x_1}(0)\) and \(p_{x_k}(0) = 1\) for each \(k \geq 2\), we have \(p\{x_1 + x_2 + \cdots = 1\} = \frac{1}{2}\). However, the Hewitt-Savage 0-1 Law is not a first-best theorem; it is still possible to relax the “identical distribution” hypothesis to a certain extent. The following exercises outline two such generalizations of this 0-1 law.
Exercise 5.9. (Horn-Schach) Let \( Y \) be a separable metric space and \( (x_m) \) a sequence of independent \( Y \)-valued random variables on a probability space \( (X, \Sigma, p) \). Assume that for each \( k \in \mathbb{N} \) there is an \( m \in \mathbb{N} \) such that \( p_{x_k} = p_{x_{k+m}} \). Prove that \( p(S) \in \{0, 1\} \) for every \( S \in \mathcal{E}(x_m) \).

Exercise 5.10. (Blum-Pathak) Let \( Y \) be a separable metric space and \( (x_m) \) a sequence of independent \( Y \)-valued random variables on a probability space \( (X, \Sigma, p) \). Assume that for each \( \varepsilon > 0 \) and \( (k, m) \in \mathbb{N}^2 \), there is an \( n \geq m \) such that \( \sup \{|p_{x_k}(B) - p_{x_n}(B)| : B \in \mathcal{B}(Y)\} < \varepsilon \). Prove that \( p(S) \in \{0, 1\} \) for every \( S \in \mathcal{E}(x_m) \).

5.3 Application: Random Walks

We now introduce a particular stochastic process which, despite its apparently simple structure, is of great importance for probability theory.

**Definition.** By a (one-dimensional) **random walk**, we mean a sequence \( (y_m) \) of random variables on a probability space \( (X, \Sigma, p) \) such that

\[
y_m = x_1 + \cdots + x_m, \quad m = 1, 2, \ldots
\]

for some sequence \( (x_m) \) of i.i.d. random variables on \( (X, \Sigma, p) \). (Here \( x_m \) is referred to as the \( m \)th step, and \( (x_m) \) as the **step sequence**, of \( (y_m) \).) If the distribution of \( x_1 \) (hence of any \( x_m \) here) is symmetric about 0, we say that \( (y_m) \) is **symmetric**. If each \( x_m \) is \( \mathbb{Z} \)-valued, then \( (y_m) \) is said to be a **random walk on \( \mathbb{Z} \)**.\(^{30}\)

In words, a random walk is the sequence of partial sums of a sequence of i.i.d. random variables. Intuitively, we may think of a random walk \( (y_m) \) as a stochastic process (through discrete time) of the following nature. We start at 0. At time 1, we take a step (of a certain random size) to the right or to the left according to a given probability distribution. We model this step by a random variable \( x_1 \), the interpretation being we go to the right of zero if \( x_1 > 0 \) and to the left of zero if \( x_1 < 0 \). (With some probability, we may stay at zero.) Thus, at the end of time 1, we find ourselves at \( x_1 \). At time 2, the process repeats itself independently of where we actually are at time 1. That is, we take an additional step \( x_2 \) (which may be to the right \( (x_2 > 0) \) or to the left \( (x_2 < 0) \)) according to the same probability distribution and independently of what \( x_1 \) is. This puts us at \( x_1 + x_2 \) at the end of time 2. And on and on. While all this is random, the structure of the process is very rigid: every one of our steps are drawn from the same distribution and independently of where

\(^{30}\)A systematic investigation of random walks was undertaken first by George Pólya in 1912. (Pólya himself coined the term “random walk” in a 1921 paper.) Apparently, Pólya’s initial interest was determining the likelihood of a multi-dimensional random walk to visit a state over and over again. Durrett (2010, p. 190) puts this rather colorfully: “Legend has it that Pólya thought of this problem while wondering around a park near Zürich when he noticed that he kept encountering the same young couple. History does not record what the young couple thought.” For more on this story, and much more on George Pólya, a great man in many accounts, have a look at the very enjoyable account sketched in Alexanderson (2000).
these steps are taken. (That is, \((x_m)\) is i.i.d.) It is “as if” we are modeling the walk of a one-dimensional creature who is so drunk that every one of its steps exhibits the same randomness no matter what and where that step is taken; this creature does not sober up even a little bit as time goes by!

**Warning.** We are a far cry from considering random walks in full generality. A more complete treatment would allow for each \(x_m\) to be at least \(\mathbb{R}^n\)-valued in the definition of above. Indeed, the theory of multi-dimensional random walks is a highly developed and an exciting topic. In this introductory text, however, we will have to limit our scope to one-dimensional random walks alone.

**Example 5.4.** Let \((y_m)\) be a sequence of random variables on a probability space such that \(y_m =_{a.s.} 0\) for every \(m\). Then \((y_m)\) is a random walk; we refer to it as the **degenerate random walk**. □

**Example 5.5.** Let \((x_m)\) be a sequence of i.i.d. random variables such that each \(x_m\) is \([-1, 1]\)-valued. Then, \((\sum_{i \in [m]} x_i)\) is a random walk on \(\mathbb{Z}\); it is said to be a **simple random walk**. The \(m\)th term of this random walk has the Binomial distribution with parameters \(m\) and \(p_{x_1}\{1\}\). In particular, this random walk is symmetric iff \(x_1\) (hence each \(x_m\)) equals \(-1\) with probability \(\frac{1}{2}\) and \(1\) with probability \(\frac{1}{2}\). □

As they provide a natural entry point to the modern theory of stochastic processes, random walks are of great importance for probability theory at large. Besides, there are a variety of real-world contexts in which random walks are indispensable tools of analysis. For instance, a major debate in the theory of finance is if the through-time behavior of stock prices exhibit the (realization) of a random walk. (This would mean that the predictable component of stock market prices are close to nil!)\(^{31}\) Similarly, there is some evidence to the effect that the shots of a basketball player (or team) through games can be modeled as a simple random walk. (If this were true, then the phrase “he has the hot hand,” which is often used by the basketball commentators, would not mean much.)

We shall touch upon various aspects of the theory of (one-dimensional) random walks in the remainder of this text by way of applying probabilistic results that we obtain in more general contexts. The following is but an illustration of this. We characterize, by using the Hewitt-Savage 0-1 Law, the asymptotic structure of all random walks with probability one.

**Lemma 5.2.** For any random walk \((y_m)\), both \(\limsup y_m\) and \(\liminf y_m\) are almost surely constant \(\mathbb{R}\)-valued random variables.

\(^{31}\)This debate was started by Burton Malkiel in 1973 – the title of Malkiel’s famous book is *A Random Walk Down Wall Street* – and is still active today; see, for instance, Lo and Mackinlay (2002).
**Proof.** Define $z := \limsup y_m$, and note that $z$ is an $\mathbb{R}$-valued random variable on the probability space, say, $(X, \Sigma, p)$, underlying the random walk $(y_m)$. We set

$$a := \inf \{ \alpha \in \mathbb{R} : p\{ z \leq \alpha \} = 1 \}.$$

If $a = -\infty$, then $z = \text{a.s. } -\infty$, so we are done; we thus assume that $a > -\infty$. Then, $p\{ z \leq a - \frac{1}{k} \} < 1$ for each $k \in \mathbb{N}$. But, by definition of a random walk, $\{ z \leq a - \frac{1}{k} \}$ is an exchangeable event associated with the step sequence of $(y_m)$ (Example 5.3). Thus, by the Hewitt-Savage 0-1 Law, $p\{ z \leq a - \frac{1}{k} \} = 0$ for each $k \in \mathbb{N}$. It follows that $p\{ z < a \} = 0$ by continuity of $p$. On the other hand, clearly, $p\{ z \leq a + \frac{1}{k} \} = 1$ for each $k \in \mathbb{N}$, so we have $p\{ z \leq a \} = 1$ again by continuity of $p$. Conclusion: $z = \text{a.s. } a$.

As the analogous analysis apply also to $\liminf y_m$, we are done.

This result does not tell the whole story. In fact, the limsup (or the liminf) of a random walk is a real number iff that random walk is degenerate (and hence it does not move at all). More precisely:

**Proposition 5.3.** For any non-degenerate random walk $(y_m)$, either $y_m \rightarrow \text{a.s. } \infty$, or $y_m \rightarrow \text{a.s. } -\infty$, or $\limsup y_m = \text{a.s. } \infty$ and $\liminf y_m = \text{a.s. } -\infty$.

**Proof.** By Lemma 5.2, $\limsup y_m = \text{a.s. } a$ for some extended real number $a$. Define $z_m := y_{m+1} - y_1$ for each $m = 1, 2, \ldots$. Notice that the distributions of $y_m$ and $z_m$ are the same, so we have $\limsup z_m = \text{a.s. } a$ as well. It follows that $a = \text{a.s. } a - y_1$. Thus, if $a$ is a real number, we have $y_1 = \text{a.s. } 0$, that is, $(y_m)$ is degenerate. Conclusion: $a \in \{ -\infty, \infty \}$. As the analogous reasoning also applies to $\liminf y_m$, we see that $\limsup y_m$ and $\liminf y_m$ are $\{ -\infty, \infty \}$-valued random variables. Our assertion follows readily from this observation.