Chapter D
Expectation via the Lebesgue Integral

Our work in the previous chapter puts us in an ideal position to introduce the expectation functional as the Lebesgue integral within the context of probability spaces. A good part of the present chapter is devoted to doing precisely this. Once the basic properties of the expectation functional are identified, we then extend the coverage of Lebesgue integration first to finite, and then to \( \sigma \)-finite, measure spaces. We also give the relevant definitions, and some very basic results, in the context of an arbitrary measure space as well, but devote relatively little time to that context.

The earlier part of our treatment is quite leisurely. In particular, we take the monotone convergence theorem as the principal step of the development, and introduce this result and its consequences (such as the linearity of the expectation functional) by means of a step-by-step approach. We then investigate the basic properties of the Lebesgue integral, introduce the notion of absolute continuity (for functions and measures), study topics like change of variables and uniform integrability, and derive some elementary probability inequalities, such as those of Jensen, Markov, and Chebyshev-Bienaymé. The primary focus of the chapter is on the integration of \( \mathbb{R} \)-valued random variables, but in an optional section, we also discuss how one may think about the expectation of Banach space-valued random variables (by means of the Bochner integral). Finally, by way of an application, we provide an introduction to the (first-order) stochastic dominance theory.

The rest of the chapter is a bit more advanced. In this part, we first sketch an introduction to the Banach spaces of integrable random variables, the so-called \( L^p \)-spaces. After discussing the basic properties of these spaces, we prove some approximation-by-continuity type theorems, and establish the connections between \( L^p \)-boundedness and uniform integrability. Finally, we present a fairly general account of the all-important Riesz-Radon Representation Theorem, and use this result to establish Choquet’s Theorem, a celebrated crossbreed between probability theory and convex analysis.\(^1\)

\(^1\)All of the texts I mentioned at the beginning of Chapter B cover most of the material we go through in this chapter with varying degrees of emphasis. There are a few exceptions, however. Probability theory books rarely cover topics such as stochastic dominance theory, integration of Banach space-valued random variables, Riesz-Radon Representation Theorem, and Choquet’s Theorem. I will suggest further references about these in due course.
1 The Expectation Functional I

The expected value of a simple random variable is the weighted average of the values taken by this random variable, where the weight of each value is its probability. You might recall that, by using (Riemann) integration, we can extend this idea to the case of a random variable whose distribution is induced by a density function. By contrast, we wish to consider here the situation for an arbitrary random variable.

1.1 Expectation of Simple Random Variables

Expectation as a Weighted Average

Let \((X, \Sigma, p)\) be a probability space. Recall that the set of all simple random variables on \((X, \Sigma, p)\) is a linear space relative to the pointwise defined addition and scalar multiplication operations (Remark B.5.3). We will define the expectation functional on this linear space first.

This is an easy task. By definition, for any simple random variable \(x\) on \((X, \Sigma, p)\), we have \(|x(X)| < \infty\) and

\[
x = \sum_{a \in x(X)} a 1_{\{x = a\}},
\]

where, as usual, \(1_{\{x = a\}}\) stands for the indicator function of the event \(\{x = a\} \in \Sigma\) on \(X\). We define the expectation of any such \(x\) as the real number

\[
\mathbb{E}(x) := \sum_{a \in x(X)} ap_{x}(a).
\]

Thus, the expected value of \(x\) is the weighted average of its values, where the weight of \(a\) is \(p_{x}(a)\) for each \(a \in x(X)\). That is, \(\mathbb{E}(x) = \sum_{a \in x(X)} ap_{x}(a)\).

**Notation.** The expectation of a simple random variable depends on the underlying probability measure \(p\), so a notation like \(\mathbb{E}_{p}(x)\) would be more informative than \(\mathbb{E}(x)\). Indeed, we will use the latter notation in what follows when we need to be specific about the probability measure we use for the computation. But when there is only one probability measure in play, we prefer to use the simpler notation \(\mathbb{E}(x)\).

The following exercise provides an easy way of computing the expectation of a simple random variable which may have been expressed in a way different than (1). We will make use of this exercise shortly.

**Exercise 1.1.** Let \((X, \Sigma, p)\) be a probability space, \(\mathcal{S} \subseteq \Sigma\) a finite partition of \(X\), and \(\varphi \in \mathbb{R}^{\mathcal{S}}\). Let \(x\) be the simple random variable on \((X, \Sigma, p)\) with \(x|_{\mathcal{S}} = \varphi(S)\) for each \(S \in \mathcal{S}\). (That is, \(x = \sum_{S \in \mathcal{S}} \varphi(S) 1_{S}\).) Show that

\[
\mathbb{E}(x) = \sum_{S \in \mathcal{S}} \varphi(S)p(S).
\]
**Linearity of $\mathbb{E}$**

Let us establish some of the basic properties of the map $\mathbb{E}$. Our first task is to show that $\mathbb{E}$ behaves linearly on its entire domain. (This observation, albeit simple, is one of the major building blocks of the subsequent development). Indeed, it is plain from our definition that this map is *homogeneous*, that is, $\mathbb{E}(ax) = a\mathbb{E}(x)$ for any real number $a$ and simple random variable $x$ on $(X, \Sigma, p)$. To see that $\mathbb{E}$ is also *additive*, take any two simple random variables $x$ and $y$ on $(X, \Sigma, p)$; and put $x = \mathbf{1}_{A(a,b)}$ and $y = \mathbf{1}_{A(a,b)}$ for each $(a, b) \in \Lambda$. Clearly, $\{A(a,b) : (a, b) \in \Lambda\}$ is a finite partition of $X$, and hence the following identities:

$$x = \sum_{(a,b)\in\Lambda} a\mathbf{1}_{A(a,b)}, \quad y = \sum_{(a,b)\in\Lambda} b\mathbf{1}_{A(a,b)}, \quad \text{and} \quad x + y = \sum_{(a,b)\in\Lambda} (a + b)\mathbf{1}_{A(a,b)}.$$ 

Thus, applying the observation noted in Exercise 1.1 three times, we find

$$\mathbb{E}(x + y) = \sum_{(a,b)\in\Lambda} (a + b)p(A(a,b)) = \sum_{(a,b)\in\Lambda} ap(A(a,b)) + \sum_{(a,b)\in\Lambda} bp(A(a,b)) = \mathbb{E}(x) + \mathbb{E}(y),$$

as we sought.

We should also emphasize that using the additivity of $\mathbb{E}$ inductively yields

$$\mathbb{E}\left(\sum_{i\in[n]} x_i\right) = \sum_{i\in[n]} \mathbb{E}(x_i) \quad (3)$$

for any positive integer $n$ and simple random variables $x_1, ..., x_n$ on $(X, \Sigma, p)$. We will henceforth use this property routinely.

**Monotonicity of $\mathbb{E}$**

Endowing the linear space of all simple random variables on a given probability space $(X, \Sigma, p)$ with the preorder $\succeq_{a.s.}$ yields a preordered linear space. Moreover, it follows readily from its definition that $\mathbb{E}$ is a positive functional on this space, that is, $x \succeq_{a.s.} 0$ implies $\mathbb{E}(x) \geq 0$ for any simple random variable $x$ on $(X, \Sigma, p)$. This implies that $\mathbb{E}$ is increasing with respect to the preorder $\succeq_{a.s.}$. Indeed, suppose $x$ and $y$ are two simple random variables on $(X, \Sigma, p)$ with $x \succeq_{a.s.} y$. Then, $x - y \succeq_{a.s.} 0$, and hence $\mathbb{E}(x - y) \geq 0$ by positivity of $\mathbb{E}$. But, by linearity of $\mathbb{E}$, we have $\mathbb{E}(x) - \mathbb{E}(y) = \mathbb{E}(x - y)$, and hence $\mathbb{E}(x) \geq \mathbb{E}(y)$. We proved:

---

2Take a quick look at Section 1.4 of Appendix 3 if this terminology is not familiar to you.
\[ \mathbb{E}(x) \geq \mathbb{E}(y) \quad \text{whenever} \quad x \geq y \text{ a.s.} \]

for any simple random variables \( x \) and \( y \) on \((X, \Sigma, \mathbb{P})\). Besides, since \( x = \text{a.s.} \ y \) iff \( x \geq \text{a.s.} \ y \geq \text{a.s.} \ x \), an immediate consequence of the monotonicity of \( \mathbb{E} \) is:

\[ \mathbb{E}(x) = \mathbb{E}(y) \quad \text{whenever} \quad x = y \text{ a.s.} \]

for any simple random variables \( x \) and \( y \) on \((X, \Sigma, \mathbb{P})\). In what follows we will use these properties of the expectation functional freely.

**Insight.** The set of all simple random variables on a probability space \((X, \Sigma, \mathbb{P})\) is a preordered linear space (relative to the pointwise defined addition and scalar multiplication operations, and the preorder \( \geq \text{a.s.} \)). The map \( \mathbb{E} \) acts as an increasing linear functional on this space.

**Variance of Simple Random Variables**

The **variance** of a simple random variable \( x \) on \((X, \Sigma, \mathbb{P})\) is the number

\[ \mathbb{V}(x) := \mathbb{E}((x - \mathbb{E}(x))^2), \]

where \( (x - \mathbb{E}(x))^2 \) is the simple random variable \( \omega \mapsto (x(\omega) - \mathbb{E}(x))^2 \) on \((X, \Sigma, \mathbb{P})\). However, it is often more convenient to use the alternate formula

\[ \mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \]

when computing the variance of \( x \). (This formula is obtained easily from the definition of \( \mathbb{V}(x) \) by using the linearity of \( \mathbb{E} \).)

**First Examples**

**Example 1.1.** Let \( n \) be a positive integer and \( p \) a number in \([0, 1]\). A \( \{0, \ldots, n\} \)-valued random variable \( x \) on a probability space \((X, \Sigma, \mathbb{P})\) such that

\[ p\{x = i\} = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, \ldots, n, \]

is said to have a binomial distribution with parameters \( n \) and \( p \). If \( n = 1 \) here, we say that \( x \) has a Bernoulli distribution with parameter \( p \).

As it is a simple random variable, we can compute the expectation of \( x \) straight from the definition. In particular, when \( n = 1 \), we have \( \mathbb{E}(x) = p(1) + (1 - p)(0) = p \), that is, the expected value of any random variable that has a Bernoulli distribution with parameter \( p \) is \( p \). In this case, we also find that \( \mathbb{E}(x^2) = p \), so \( \mathbb{V}(x) = p - p^2 \).
More generally, for any random variable that has a binomial distribution with parameters \( n \) and \( p \), (2) takes the form

\[
E(x) = \sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i}.
\]

To compute this sum, let us first note that \( n\binom{n-1}{i-1} = \binom{n}{i} \) for any \( n \in \mathbb{N} \) and \( i \in [n] \). Consequently,

\[
E(x) = np \sum_{i \in [n]} \binom{n-1}{i-1} p^{i-1} (1 - p)^{n-i} = np \sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1 - p)^{(n-1)-i},
\]

so, by the Binomial Theorem, \( E(x) = np(p + (1 - p))^{n-1} = np \). By using a similar method, we can also show that \( E(x^2) = n^2 p^2 - np^2 + np \), so \( \text{V}(x) = np(1 - p) \). □

Example 1.2. Take any positive integer \( n \), and recall that a permutation on \([n]\) is a bijection from \([n]\) onto \([n]\). Question: How many numbers in \([n]\) do you expect that a random permutation would leave fixed?

Let us first formalize the problem. Clearly, there are \( n! \) many distinct permutations on \([n]\). Thus, \((X, 2^X, \mathbf{p})\), where \( X \) is the set of all permutations on \([n]\) and \( \mathbf{p} : 2^X \rightarrow [0, 1] \) satisfies \( \mathbf{p} \{ \omega \} := 1/n! \), is a probability space. Define \( x : X \rightarrow [n] \) as the map

\[
x(\omega) := |\{ i \in [n] : \omega(i) = i \}|.
\]

Our problem is to determine \( E(x) \). To this end, for each \( i \in [n] \), we consider the map \( x_i : X \rightarrow \{0, 1\} \) with \( x_i(\omega) = 1 \) iff \( \omega(i) = i \). As there are exactly \((n - 1)! \) many permutations on \([n]\) that map \( i \) to \( i \), we have \( E(x_i) = (n - 1)!/n! \), for each \( i \in [n] \). But, clearly, \( x = x_1 + \cdots + x_n \). Thus, by (3), we have \( E(x) = 1 \). Conclusion: In expectation, a random permutation of \( n \) numbers leaves exactly one number fixed. □

1.2 Expectation of Nonnegative Random Variables

Expectation by Approximation from Below

Let \( x \) be a \([0, \infty]\)-valued random variable on a probability space \((X, \Sigma, \mathbf{p})\). We define the expectation of \( x \) as the (extended real) number

\[
E(x) := \sup \{ E(z) : z \in \mathcal{L}(x) \}
\]

where \( \mathcal{L}(x) \) stands for the set of all simple random variables \( z \) such that \( z \leq x \). In turn, the variance of \( x \) is defined as the (extended real) number

\[
\text{V}(x) := E((x - E(x))^2),
\]

5
provided that $E(x) < \infty$. (Here, of course, $x - E(x)$ is the map $\omega \mapsto x(\omega) - E(x)$ on $X$.) Please note that these definitions agree with those we gave in Section 1.1 in the case of simple random variables on $(X, \Sigma, p)$. Moreover, they extend those definitions to the case of $[0, \infty]$-valued simple random variables.

**FIGURE D.1 ABOUT HERE**

In words, the (extended real) number $E(x)$ is defined as the supremum of the weighted averages of all those simple random variables that are “smaller than” $x$ everywhere. So, in this sense, the idea behind the definition of $E(x)$ is reminiscent of that of the computation of the area under a given curve in $\mathbb{R} \times \mathbb{R}_+$ by approximating this area with the sum of the areas of the rectangles that lie under the curve and above the horizontal axis. (See Figure D.1.) Put this way, you should see that $E(x)$ can be thought of as some sort of an integral of $x$ – it is called the Lebesgue integral of $x$ with respect to $p$ – where the sets in the domain of $x$ (analogous to the bases of the rectangles under the curve) are “measured” according to the underlying probability measure.\(^3\)

The commonly used notation for the Lebesgue integral of $x$ on $X$ with respect to $p$ is $\int_X x dp$, that is,

$$\int_X x dp \quad \text{and} \quad E(x)$$

denote the same (extended real) number. Adopting the widely used conventions of integration theory, we also set

$$\int_S x dp := E(x 1_S) \quad \text{for any } S \in \Sigma,$$

where $1_S$ is the indicator function of $S$ on $X$. Therefore, recalling the definition of the expectation of a simple random variable, we see that $\int_S dp$, that is, the Lebesgue integral of the constant function $1$ with respect to $p$ on $S$, and $E(1_S)$ denote the same number, namely, $p(S)$, for any $S \in \Sigma$.

\(^3\)I’m allowing myself to be a bit loose here. This discussion, and especially Figure D.1, while suggestive, actually overstates the intuition. Indeed, even for a simple nonnegative random variable $z$ on a probability space like $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$, the number $E(z)$ need not correspond to the sum of the areas of some rectangles, because for any $a \in z(X)$, the event \{z = a\} need not be an interval (although it is sure to be a Borel set in $\mathbb{R}$). In fact, this is the source of the main difference between the Lebesgue integral and the more familiar Riemann integral. I will talk about this issue at length in the next chapter.

On a related note, you may ask yourself why I defined $E$ by approximation from below instead of approximation from above. This is because the latter approach would run into difficulties for nonnegative random variables that are not bounded from above. For the bounded ones, however, this alternative approach would work just as well; see Exercise 1.6 below.
Notation. Many authors write

$$\int_X x(\omega)p(d\omega)$$

instead of $$\int_X xd\mathbf{p}$$. The Lebesgue integral of the map $$\omega \mapsto \omega^2$$ on the probability space $$([0,1], \mathcal{B}[0,1], \ell)$$ is, for instance, written as $$\int_{[0,1]} \omega^2 \ell(d\omega)$$. We shall adopt this notation on occasion as well.

Monotonicity of $$\mathbb{E}$$, Again

Our present definition of $$\mathbb{E}$$ applies to a much larger collection of random variables than that of simple random variables. Nevertheless, this definition inherits the fundamental properties of our earlier, more restricted definition. In particular, $$\mathbb{E}$$ remains monotonic. Indeed, for any $$[0, \infty]$$-valued random variables $$x$$ and $$y$$ on a probability space $$(X, \Sigma, \mathbf{p})$$ such that $$y \geq x$$, we clearly have $$\mathcal{L}(y) \supseteq \mathcal{L}(x)$$, and hence $$\mathbb{E}(y) \geq \mathbb{E}(x)$$.

We can actually do better than this, because, in fact,

$$\mathbb{E}(x) = \sup\{\mathbb{E}(z) : z \in \mathcal{L}^*(x)\},$$

where $$\mathcal{L}^*(x)$$ is the set of all simple random variables $$z$$ such that $$z \leq_{a.s.} x$$, and similarly for $$\mathbb{E}(y)$$. To see this, take any $$z \in \mathcal{L}^*(x)$$, put $$S := \{z \leq x\}$$, and note that $$S \in \Sigma$$. Then, $$z1_S$$ is a simple random variable on $$(X, \Sigma, \mathbf{p})$$ with $$z1_S \leq x$$, that is, $$z1_S \in \mathcal{L}(x)$$. But $$z$$ and $$z1_S$$ are simple random variables that are almost surely equal to each other, so $$\mathbb{E}(z) = \mathbb{E}(z1_S)$$. It follows that $$\mathbb{E}(x) \geq \sup\{\mathbb{E}(z) : z \in \mathcal{L}^*(x)\}$$. Besides, since $$\mathcal{L}(x) \subseteq \mathcal{L}^*(x)$$, the converse of this inequality is trivially true.

Now, a moment’s reflection shows that we have $$\mathcal{L}^*(y) \subseteq \mathcal{L}^*(x)$$ whenever $$x \geq_{a.s.} y$$. But then, in view of (4) and the analogous expression for $$\mathbb{E}(y)$$, we find that $$\mathbb{E}(x) \geq \mathbb{E}(y)$$ whenever $$x \geq_{a.s.} y$$. This is what we mean by the monotonicity of $$\mathbb{E}$$.

**Proposition 1.1.** For any $$[0, \infty]$$-valued random variables $$x$$ and $$y$$ (on a given probability space),

$$x \geq_{a.s.} y \implies \mathbb{E}(x) \geq \mathbb{E}(y),$$

and hence

$$x =_{a.s.} y \implies \mathbb{E}(x) = \mathbb{E}(y).$$

In the following set of exercises $$x$$ stands for an arbitrary $$[0, \infty]$$-valued random variable on a probability space $$(X, \Sigma, \mathbf{p})$$.  

**Exercise 1.2.** For any real number $$a \geq 0$$, and $$A, B \in \Sigma$$ with $$A \supseteq B$$, prove that

$$\int_X axd\mathbf{p} = a \int_X xd\mathbf{p} \quad \text{and} \quad \int_A xd\mathbf{p} \geq \int_B xd\mathbf{p}.$$

**Exercise 1.3.** For any $$m, n \in \mathbb{N}$$ such that $$m < n$$ and $$\mathbb{E}(x^n) < \infty$$, show that $$\mathbb{E}(x^m) < \infty$$.  

7
Exercise 1.4. Assume $x > 0$, and show that
\[ \int_{\{x > m\}} \frac{m}{x} \, dp \to 0 \quad \text{as } m \uparrow \infty. \]

Exercise 1.5. Prove: $x = \text{a.s. } 0$ iff $\mathbb{E}(x) = 0$.

Exercise 1.6. Assume that $x$ is bounded (that is, $\|x\|_\infty < \infty$), and prove:
\[ \mathbb{E}(x) = \inf \{ \mathbb{E}(z) : z \text{ is a simple random variable on } (X, \Sigma, p) \text{ with } z \geq x \}. \]

**Monotone Convergence Theorem**

Our next task is to check for the *additivity* of $\mathbb{E}$. Unfortunately, things are less trivial on this front. In fact, we will launch a rather indirect attack on this problem, by first proving a fundamental convergence theorem for $\mathbb{E}$. An equivalent version of this result was first proved by Beppo Levi in 1906.

**The Monotone Convergence Theorem 1.** Let $x, x_1, x_2, \ldots$ be $[0, \infty]$-valued random variables on a probability space $(X, \Sigma, p)$. Then,
\[ x_m \uparrow \text{a.s. } x \quad \text{implies} \quad \mathbb{E}(x_m) \uparrow \mathbb{E}(x). \]

**Proof.** By Proposition 1.1, $\mathbb{E}(x_1) \leq \mathbb{E}(x_2) \leq \cdots \leq \mathbb{E}(x)$, so all we need to do here is to establish that $\sup \mathbb{E}(x_m) \geq \mathbb{E}(x)$.\(^4\) Recalling the definition of $\mathbb{E}(x)$, our proof will then be complete if we can show that $\sup \mathbb{E}(x_m) \geq \mathbb{E}(z)$ for every $z \in \mathcal{L}(x)$.

To prove this, fix an arbitrary $z \in \mathcal{L}(x)$ and $\varepsilon > 0$, and define $A_m := \{ x_m < z - \varepsilon \}$ for every positive integer $m$.\(^5\) (We have $A_m \in \Sigma$ for each $m$; why?) Since
\[ z = (z - \varepsilon) 1_{X \setminus A_m} + z 1_{A_m} + \varepsilon 1_{X \setminus A_m}, \]
and $\mathbb{E}$ is a linear functional over the linear space of simple random variables on $(X, \Sigma, p)$,
\[
\mathbb{E}(x_m) - \mathbb{E}(z) = \mathbb{E}(x_m) - \int_{X \setminus A_m} (z - \varepsilon) \, dp - \int_{A_m} zd \, p - \varepsilon \int_{X \setminus A_m} d \, p \geq \left( \mathbb{E}(x_m) - \int_{X \setminus A_m} (z - \varepsilon) \, d \, p \right) - \|z\|_\infty p(A_m) - \varepsilon p(X \setminus A_m).
\]

\(^4\)Here, to simplify the notation, I write $\sup \mathbb{E}(x_m)$ for the number $\sup \{ \mathbb{E}(x_m) : m \in \mathbb{N} \}$.

\(^5\)Observe that the only reason why I can’t safely conclude that $\mathbb{E}(x_M) \geq \mathbb{E}(z) - \varepsilon$ for some $M$ (and hence for all $m \geq M$) is that $z - \varepsilon$ is larger than $x_M$ on the set $A_M$. Indeed, if $A_M = \emptyset$ for some $M$, then it follows readily that $\mathbb{E}(x_m) \geq \mathbb{E}(z) - \varepsilon$ for each $m \geq M$ (since $A_1 \supseteq A_2 \supseteq \cdots$). Unfortunately, I don’t know if $A_M = \emptyset$ for some $M$. But I do know that $A_m$s can be ignored in the limit, for $\bigcap_1^\infty A_i = \emptyset$ so that $p(A_m) \downarrow 0$. So, for $m$ large enough, the problematic region $A_m$ should not really matter. It is this intuition that the proof is built on.
for each \( m \). But by definition of \( A_m \), we have \( x_m \geq (z - \varepsilon)1_{X \setminus A_m} \), so that
\[
\mathbb{E}(x_m) - \int_{X \setminus A_m} (z - \varepsilon) \, dp = \mathbb{E}(x_m) - \mathbb{E}
\left((z - \varepsilon)1_{X \setminus A_m}\right) \geq 0
\]
for each \( m \). Combining this inequality with the previous one, and using the fact that \( p(X \setminus A_m) \leq 1 \), we obtain
\[
\mathbb{E}(x_m) - \mathbb{E}(z) \geq -\|z\|_{\infty} p(A_m) - \varepsilon, \quad m = 1, 2, \ldots \tag{5}
\]
Now, let us assume for a moment that we actually have \( x_m \uparrow x \). Then, \( x_1 \leq x_2 \leq \cdots \), and this guarantees that \( (A_m) \) is a decreasing sequence. Moreover, since \( z \leq x \) and \( x_m \to x \), it must be the case that, for any \( \omega \in X \), we have \( x_m(\omega) > z(\omega) - \varepsilon \) for \( m \) large enough. Thus: \( \bigcap_{i=1}^{\infty} A_i = \emptyset \). So, by continuity of \( p \), we find \( p(A_m) \downarrow 0 \). Therefore, letting \( m \uparrow \infty \) in (5), we find \( \sup \mathbb{E}(x_m) = \lim \mathbb{E}(x_m) \geq \mathbb{E}(z) - \varepsilon \). Since \( \varepsilon > 0 \) is arbitrary here, we are done.

To complete the proof, define \( S := \{x_m \uparrow x\} \), and note that \( x1_S =_{a.s.} x \) and \( x_m 1_S =_{a.s.} x_m \) for each \( m \).\footnote{We know from Section C.2.3 that \( S \) belongs to \( \Sigma \). As the product of two random variables is itself a random variable, therefore, each \( x_m 1_S \) (and also \( x 1_S \)) is a \([0, \infty]\)-valued random variable on \((X, \Sigma, p)\).} It follows from Proposition 1.1 that \( \mathbb{E}(x1_S) = \mathbb{E}(x) \) and \( \mathbb{E}(x_m 1_S) = \mathbb{E}(x_m) \) for each \( m \). Furthermore, we have \( x_m 1_S \uparrow x 1_S \), so, by what we have just proved, \( \mathbb{E}(x_m 1_S) \uparrow \mathbb{E}(x 1_S) \), and hence \( \mathbb{E}(x_m) \uparrow \mathbb{E}(x) \), as we sought.

The “trick” we used in the last paragraph of the proof above is frequently useful when working with almost surely convergent sequences of random variables. The idea is simply to replace the involved random variables with suitable “dummy” random variables so that we obtain an (everywhere) convergent sequence whose \( i \)th term equals that of the original sequence almost surely. Most probabilistic results we would obtain in terms of this “dummy” sequence may then easily be expressed also in terms of our original sequence. We will use this trick again and again in this chapter.

Here is a first application of the Monotone Convergence Theorem 1.

**Example 1.3.** Let \((X, 2^X, p)\) be a probability space with \( X \) being a countable set. We wish to find an expression for the expectation of an arbitrary nonnegative random variable \( x \) on \((X, 2^X, p)\).

We know the answer if \( X \) is finite, for then \( x \) is simple by necessity, and we have an explicit formula for the expectation of simple random variables. To focus on the nontrivial case, then, we assume that \( X \) is countably infinite, and enumerate it as \( X := \{\omega_1, \omega_2, \ldots\} \). For every positive integer \( m \), let us define the simple random variable
\[
x_m(\omega) := \begin{cases} x(\omega), & \text{if } \omega \in \{\omega_1, \ldots, \omega_m\} \\ 0, & \text{otherwise.} \end{cases}
\]
By definition of $E$ for simple random variables, we have $E(x_m) = \sum_{i \in [m]} x(\omega_i)p(\omega_i)$ for each $m$. But, $x_m \uparrow x$, so we have $E(x_m) \uparrow E(x)$ by the Monotone Convergence Theorem 1. Consequently,

$$E(x) = \lim E(x_m) = \lim \sum_{i \in [m]} x(\omega_i)p(\omega_i) = \sum_{i=1}^{\infty} x(\omega_i)p(\omega_i).$$

In view of Proposition 1.8 of Appendix 1, the number on the rightmost side of this statement is independent of the way we have enumerated $X$, so we may write

$$E(x) = \sum_{\omega \in X} x(\omega)p(\omega)$$

for any nonnegative random variable $x$ on $(X, 2^X, p).$  

A Non-Monotone Convergence-Like Theorem

The way it is stated, the Monotone Convergence Theorem 1 is useful only when establishing the convergence of integrals of increasing sequences of nonnegative random variables. The following celebrated result, which was proved by Pierre Fatou in 1906, often proves useful when studying the convergence problem for arbitrary sequences of nonnegative random variables.

**Fatou’s Lemma.** Let $x, x_1, x_2, \ldots$ be $[0, \infty]$-valued random variables on a probability space $(X, \Sigma, p)$ such that $x = \text{a.s.} \lim \inf x_m$. Then,

$$\int_X x \, dp \leq \lim \inf \left( \int_X x_m \, dp \right).$$

**Proof.** Define $y_m := \inf\{x_m, x_{m+1}, \ldots\}$ for each $m$, and note that each $y_m$, as well as $\lim y_m$, are $[0, \infty]$-valued random variables on $(X, \Sigma, p)$. (Recall Exercise C.1.8.) Moreover, we have $y_m \uparrow \lim \inf x_m$ so, by Proposition 1.1 and the Monotone Convergence Theorem 1,

$$\int_X x \, dp = \int_X \lim \inf x_m \, dp = \lim \int_X y_m \, dp \leq \lim \inf \left( \int_X x_m \, dp \right)$$

where the inequality follows from the fact that $y_m \leq x_m$ for each $m$, and Proposition 1.1.

**Warning.** The inequality in Fatou’s Lemma may hold strictly. For instance, let $(x_m)$ be the sequence of random variables defined on the probability space $([0, 1], \mathcal{B}[0, 1], \ell)$ by $x_m := m1_{(0,1/m)}$. For this sequence, we have $\lim \inf x_m = 0$ with $E(x_m) = 1$ for each $m$, so $E(\lim \inf x_m) < \lim \inf E(x_m)$. This example also shows that the monotonicity requirement of the Monotone Convergence Theorem 1 cannot be relaxed completely.
Exercise 1.7.1 (A Dominated Convergence Theorem for $[0, \infty]$-Valued Random Variables)

Let $x, x_1, x_2, \ldots$ be $[0, \infty]$-valued random variables on a probability space $(X, \Sigma, p)$ such that $x_m \to x$ and $x \geq \sup\{x_1, x_2, \ldots\}$. Prove that $E(x_m) \to E(x)$.

Linearity of $E$

We are still to prove the additivity of the map $E$ on its domain. The standard way of doing this is to use a two-step procedure. First, we approximate the given $[0, \infty]$-valued random variables by increasing sequences of simple random variables. And then, we deduce the sought additivity by using the additivity of $E$ over the set of simple random variables along with the Monotone Convergence Theorem 1.

**Proposition 1.2.** Let $x$ and $y$ be two $[0, \infty]$-valued random variables on a probability space $(X, \Sigma, p)$. Then, for any $a \geq 0$,

$$E(ax + y) = aE(x) + E(y).$$

**Proof.** In view of the first part of Exercise 1.2, it is enough to prove this claim for $a = 1$. To this end, we use Proposition C.1.10 to find increasing sequences $(x_m)$ and $(y_m)$ of nonnegative simple random variables on $(X, \Sigma, p)$ such that $x_m \uparrow x$ and $y_m \uparrow y$. Clearly, $(x_m + y_m)$ is an increasing sequence of nonnegative simple random variables such that $x_m + y_m \uparrow x + y$. So, by applying the Monotone Convergence Theorem 1 three times, we get

$$E(\lim(x_m + y_m)) = \lim E(x_m + y_m) = \lim E(x_m) + \lim E(y_m) = E(\lim x_m) + E(\lim y_m).$$

It follows that $E(x + y) = E(x) + E(y)$, as we sought. 

The additivity of $E$ is sometimes used implicitly through a partition of the sample space. For instance, for any given $[0, \infty]$-valued random variable $x$ on a probability space $(X, \Sigma, p)$, we have

$$\int_X xd\mathbf{p} = \int_S xd\mathbf{p} + \int_{X\setminus S} xd\mathbf{p}$$

for any $S \in \Sigma$. This is a (very) special case of Proposition 1.2. After all, this equation says simply that $E(x 1_S + x 1_{X\setminus S}) = E(x 1_S) + E(x 1_{X\setminus S})$ for every $S \in \Sigma$. Moreover, as you are asked to prove below, one can use a suitable monotone convergence argument to extend this observation to the case of countable partitions of $X$.

In the following set of exercises the expectations are taken relative to an arbitrarily fixed probability space $(X, \Sigma, p)$. 

11
Exercise 1.8. For any nonempty countable set \( X \) of \([0, \infty]\)-valued random variables on \((X, \Sigma, \mathbf{p})\), show that
\[
\mathbb{E} \left( \sum_{x \in \mathcal{X}} x \right) = \sum_{x \in \mathcal{X}} \mathbb{E} (x).
\]

Exercise 1.9. Let \( z \) be a \([0, \infty]\)-valued random variable on \((X, \Sigma, \mathbf{p})\), and \( \mathcal{S} \) a countable partition of \( X \) with \( \mathcal{S} \subseteq \Sigma \). Prove that
\[
\int_X x \, d\mathbf{p} = \sum_{S \in \mathcal{S}} \int_S x \, d\mathbf{p}.
\]

Exercise 1.10. Assume that \( X = [0,1] \), and show that
\[
\int_X \frac{m x(\omega)}{m + \omega} \mathbf{p}(d\omega) \to \mathbb{E}(x) \quad \text{for every } x \in L^0_+(X, \Sigma).
\]

Exercise 1.11. Let \( a \) and \( b \) be two nonnegative real numbers with \( a < b \), and take any \([a,b]\)-valued random variable on \((X, \Sigma, \mathbf{p})\). Prove:
\[
\int_X x \, d\mathbf{p} = \int_\mathbb{R} 1_{[a,b]}(x) \, d\mathbf{p}_x.
\]

Exercise 1.12. Take any \( x \in L^0_+(X, \Sigma) \) with \( \mathbb{E}(x) = 1 \), and define \( \mathbf{q} : \Sigma \to \mathbb{R}_+ \) by \( \mathbf{q}(S) := \int_S x \, d\mathbf{p} \). Prove: \( \mathbf{q} \) is a probability measure on \( \Sigma \) such that
\[
\int_X y \, d\mathbf{q} = \int_X xy \, d\mathbf{p} \quad \text{for every } y \in L^0_+(X, \Sigma).
\]

Exercise 1.13. Let \( x \) be a random variable on \((X, \Sigma, \mathbf{p})\) such that \( x \geq 0 \) and \( \mathbb{E}(x) < \infty \). Take any \( (S_m) \in \Sigma^\infty \) with \( \lim \mathbf{p}(S_m) = 0 \). Show that \( x 1_{S_m} \to \text{a.s.} \) need not be true, but we have \( \mathbb{E}(x 1_{S_m}) \to 0 \).

Exercise 1.14. (An Axiomatic Derivation of \( \mathbb{E} \)) Let \( I : L^0_+(X, \Sigma) \to \mathbb{R}_+ \) be a function such that (i) \( I(1_X) = 1 \); (ii) \( I(ax + y) = aI(x) + I(y) \) for every \( x, y \in L^0_+(X, \Sigma) \) and \( a \geq 0 \); (iii) \( I(\lim x_m) = \lim I(x_m) \) for every increasing sequence \( (x_m) \) in \( L^0_+(X, \Sigma) \). Prove that there is a unique probability measure \( \mathbf{q} \) on \( \Sigma \) such that \( I(x) = \int_X x \, d\mathbf{q} \) for any \( x \in L^0_+(X, \Sigma) \).

1.3 The Lebesgue Integral I

We have defined the “Lebesgue integral” above only in the context of probability spaces. With minor modifications in the analysis, it is possible to extend this notion to the context of an arbitrary measure space. Moreover, we can use what we have learned in Section 1.2 to deduce the basic properties of this integral. We will outline how to do these things in this section, but as infinite measure spaces play a limited role in this text, our treatment will be concise.

Lebesgue Integration of Simple Maps
Let \((X, \Sigma, \mu)\) be a measure space. By a \textbf{nonnegative simple \(\Sigma\)-measurable map} on \(X\), we mean a real function \(f \in \mathcal{L}^0(X, \Sigma)\) such that \(f \geq 0\) and \(f(X)\) is finite. The \textbf{Lebesgue integral} of any such map with respect to \(\mu\) is defined as the number

\[
\int_X f \, d\mu := \sum_{a \in \mathcal{P}(X)} a \mu\{f = a\},
\]

where we adopt the convention that \(0 \cdot \infty\) equals \(0\). (This convention is adopted throughout this section.) Repeating the analysis we gave in Section 1.1, one can easily verify that this integral is (positively) homogeneous as well as additive. Moreover, again by the same analysis,

\[
\int_X f \, d\mu \geq \int_X g \, d\mu \quad \text{whenever} \quad \mu\{f < g\} = 0, \tag{6}
\]

and hence

\[
\int_X f \, d\mu = \int_X g \, d\mu \quad \text{whenever} \quad \mu\{f \neq g\} = 0, \tag{7}
\]

for any nonnegative simple \(\Sigma\)-measurable maps \(f\) and \(g\) on \(X\).

**Lebesgue Integration of Nonnegative Maps**

Let \((X, \Sigma, \mu)\) be a measure space. The \textbf{Lebesgue integral} of any \(\Sigma\)-measurable \(f : X \to [0, \infty]\) with respect to \(\mu\) is defined as the (extended) real number

\[
\int_X f \, d\mu := \sup \left\{ \int_X h \, d\mu : h \in \mathcal{M}(f) \right\}
\]

where \(\mathcal{M}(f)\) stands for the set of all nonnegative simple \(\Sigma\)-measurable maps \(h\) on \(X\) with \(h \leq f\). In turn, we define

\[
\int_S f \, d\mu := \int_X f 1_S \, d\mu \quad \text{for any} \ S \in \Sigma,
\]

where \(1_S\) is the indicator function of \(S\) on \(X\).

**Exercise 1.15.** Let \((X, \Sigma, \mu)\) be a measure space, and \(S \in \Sigma\). Show that

\[
\int_X f \, d\mu = \int_S f \, d\mu + \int_{X \setminus S} f \, d\mu
\]

for any \(\Sigma\)-measurable \(f : X \to [0, \infty]\).

**Exercise 1.16.** Let \((X, \Sigma, \mu)\) be a measure space, and \(S \in \Sigma \setminus \{\emptyset\}\). Put \(\Sigma_S := \{A \cap S : A \in \Sigma\}\), which is a \(\sigma\)-algebra on \(S\), and define \(\mu_S : \Sigma_S \to [0, \infty]\) by \(\mu_S(B) := \mu(B)\). Show that

\[
\int_S f \, d\mu = \int_S f 1_{S} d\mu_S
\]
for every $\Sigma$-measurable $f : X \to [0, \infty]$. (Notice that the latter integral is taken in the context of the measure space $(S, \Sigma, \mu_S)$.) Conclude from this that if $\nu$ is another measure on $\Sigma$ with $\nu(A \cap S) = \mu(A \cap S)$ for each $A \in \Sigma$, then

$$
\int_S f \, d\mu = \int_S f \, d\nu
$$

for every $\Sigma$-measurable $f : X \to [0, \infty]$.

It is plain that the Lebesgue integral is positively homogeneous, that is, $\int_X af \, d\mu = a \int_X f \, d\mu$ for any $a \geq 0$ and any $\Sigma$-measurable $f : X \to [0, \infty]$. Besides, we can repeat the argument we gave for Proposition 1.1 to establish that this integral is monotonic, that is, (6) and (7) hold for any $\Sigma$-measurable $[0, \infty]$-valued maps $f$ and $g$ on $X$ with $f \geq g$. In particular, applying (6) for $g = f \mathbf{1}_{\{f \geq a\}}$, we find

$$
\int_X f \, d\mu \geq \int_X f \mathbf{1}_{\{f \geq a\}} \, d\mu \geq \int_X a \mathbf{1}_{\{f \geq a\}} \, d\mu = a \int_X \mathbf{1}_{\{f \geq a\}} \, d\mu = \mu\{f \geq a\},
$$

that is,

$$
\mu\{f \geq a\} \leq \frac{1}{a} \int_X f \, d\mu
$$

for any $a > 0$ and $\Sigma$-measurable $f : X \to [0, \infty]$. (This is sometimes referred to as Markov’s Inequality for $\mu$.)

At this stage, we may proceed into deriving the rest of the properties of the Lebesgue integral as we did in Section 1.2 for the expectation functional, but this would at times require some nontrivial modifications (in the case of, say, the monotone convergence theorem). Fortunately, things are much simpler in the case of $\sigma$-finite measure spaces, which is the context that is immediately relevant for probabilistic analysis. As we shall see, this is because the Lebesgue integral in the context of such spaces can be additively decomposed into countably many expectation functionals.

**Lebesgue Integration on Finite Measure Spaces**

Let us first look into the particularly simple case of integrating a $\Sigma$-measurable map $f : X \to [0, \infty]$ on a finite measure space $(X, \Sigma, \mu)$. If $\mu(X) = 0$ here, all is trivial, for then $\int_X f \, d\mu = 0$. On the other hand, if $\mu(X) > 0$, then $\frac{1}{\mu(X)} f \, d\mu$ is a probability measure on $\Sigma$, and we have

$$
\int_X f \, d\mu = \mu(X) \int_X f \, d\left(\frac{1}{\mu(X)} \mu\right).
$$

Thus:

**Proposition 1.3.** Let $(X, \Sigma, \mu)$ be a finite measure space. Then, there is a real number $\lambda \geq 0$ and a probability measure $p$ on $\Sigma$ such that

$$
\int_X f \, d\mu = \lambda \int_X f \, dp
$$

14
for every \( \Sigma \)-measurable \( f : X \to [0, \infty] \).

It is plain from this observation that any one property of the Lebesgue integral on a finite measure space can be deduced from the corresponding property of the expectation functional in a straightforward manner.

**Lebesgue Integration on \( \sigma \)-Finite Measure Spaces**

Let \( (X, \Sigma, \mu) \) be a measure space. We say that this space (or \( \mu \) itself) is \( \sigma \)-finite if there is a countable partition \( S \) of \( X \) such that \( S \subseteq \Sigma \) and \( \mu(S) < \infty \) for each \( S \in S \). We may extend Proposition 1.3 to the context of such spaces as follows:

**Proposition 1.4.** Let \( (X, \Sigma, \mu) \) be a \( \sigma \)-finite measure space with \( \mu(X) = \infty \). Then, there is a sequence \((\lambda_m)\) in \( \mathbb{R}^+ \) and a sequence \((p_m)\) of probability measures on \( \Sigma \) such that

\[
\int_X f d\mu = \sum_{i=1}^{\infty} \lambda_i \int_X f d\mu_i
\]

for every \( \Sigma \)-measurable \( f : X \to [0, \infty] \).

**Proof.** Let \( S \) be a countable partition of \( X \) such that \( S \subseteq \Sigma \) and \( \mu(S) < \infty \) for each \( S \in S \). Define \( \mathcal{X} := \{ S \in \mathcal{S} : \mu(S) > 0 \} \). As \( \mu(X) = \infty \), \( \mathcal{X} \) must be a countably infinite collection, so we may enumerate it as \( \{X_1, X_2, \ldots \} \). For each positive integer \( i \), we put \( \lambda_i := \mu(X_i) \), and define \( p_i : \Sigma \to [0, 1] \) by \( p_i(T) := \frac{\mu(T \cap X_i)}{\mu(X_i)} \). Clearly, \( p_i \) is a probability measure on \( \Sigma \) for each \( i \). Moreover, for any nonnegative simple \( \Sigma \)-measurable map \( f \) on \( X \),

\[
\sum_{a \in f(X)} a \mu\{f = a\} = \sum_{a \in f(X)} \sum_{S \in S} a \mu(S \cap \{f = a\} \cap S)
= \sum_{a \in f(X)} \sum_{i=1}^{\infty} \mu(S \cap \{f = a\} \cap X_i)
= \sum_{i=1}^{\infty} \lambda_i \sum_{a \in f(X)} a p_i\{f = a\},
\]

so (8) holds.\(^7\) To complete the proof, take any \( \Sigma \)-measurable \( f : X \to [0, \infty] \). By what we have just shown,

\[
\int_X h d\mu = \sum_{i=1}^{\infty} \lambda_i \int_X h d\mu_i \leq \sum_{i=1}^{\infty} \lambda_i \int_X f d\mu_i
\]

\(^7\)The third equality is valid because \( \sum_{i=1}^{\infty} (a_i + b_i) = \sum_{i=1}^{\infty} a_i + \sum_{i=1}^{\infty} b_i \) for any sequences \((a_m)\) and \((b_m)\) of nonnegative real numbers.
for any \( h \in \mathcal{M}(f) \), and it follows that the left-hand side of (8) is less than its right-hand side. To prove the converse inequality, take any positive integer \( m \) and \( h_1, \ldots, h_m \in \mathcal{M}(f) \), and notice that

\[
\sum_{i \in [m]} \lambda_i \int_X h_i \, dp_i \leq \sum_{i \in [m]} \lambda_i \int_X \max\{h_1, \ldots, h_m\} \, dp_i \leq \int_X \max\{h_1, \ldots, h_m\} \, d\mu \leq \int_X f \, d\mu.
\]

(Why the second inequality?) Taking the supremum over all \( h_1, \ldots, h_m \) in \( \mathcal{M}(f) \) here yields

\[
\sum_{i \in [m]} \lambda_i \int_X f \, dp_i \leq \int_X f \, d\mu,
\]

and letting \( m \uparrow \infty \) completes our proof. \( \square \)

**Example 1.4.** In the case of specific \( \sigma \)-finite measure spaces, we may provide more precise formulas than (8) by choosing suitable partitions of the sample space in the proof of Proposition 1.4. For instance, consider the \( \sigma \)-finite measure space \((\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)\). If we repeat the argument in that proof with \( S \) being \( \{(i, i + 1]: i \in \mathbb{Z}\} \), we get

\[
\int_{\mathbb{R}} f \, d\ell = \sum_{i \in \mathbb{Z}} \int_{\mathbb{R}} f \, dp_i \quad \text{for every } f \in L^0_+ (\mathbb{R}, \mathcal{B}(\mathbb{R}))
\]

where \( p_i \) is the Borel probability measure on \( \mathbb{R} \) defined by \( p_i(T) := \ell(T \cap (i, i + 1]) \) for each integer \( i \). But, for any integer \( i \) and any such \( f \),

\[
\int_{\mathbb{R}} f \, dp_i = \int_{\mathbb{R}} f 1_{(i, i+1]} \, dp_i = \int_{(i, i+1]} f \, dp_i = \int_{(i, i+1]} f \, d\ell.
\]

(The first equality here is due to the fact that \( p_i \{f \neq f 1_{(i, i+1]} \} = 0 \), while the third equality follows from Exercise 1.16.) Thus:

\[
\int_{\mathbb{R}} f \, d\ell = \sum_{i \in \mathbb{Z}} \int_{(i, i+1]} f \, d\ell \quad \text{for every } f \in L^0_+ (\mathbb{R}, \mathcal{B}(\mathbb{R})).
\]

\( \square \)

**Additivity of the Lebesgue Integral**

Proposition 1.4 allows us to study the properties of the Lebesgue integral with respect to an arbitrary measure by using what we already know about the expectation functional. As a case study, let us look at the additivity property. Let \((X, \Sigma, \mu)\) be a measure space, and take any \( \Sigma \)-measurable \([0, \infty]\)-valued maps \( f \) and \( g \) on \( X \). Assume first that \((X, \Sigma, \mu)\) is \( \sigma \)-finite. Then, where \((\lambda_m)\) and \((p_m)\) are as found in
Proposition 1.4,
\[
\int_X (f + g) d\mu = \sum_{i=1}^{\infty} \lambda_i \int_X (f + g) d\mu_i
\]
\[
= \sum_{i=1}^{\infty} \lambda_i \left( \int_X f d\mu_i + \int_X g d\mu_i \right)
\]
\[
= \sum_{i=1}^{\infty} \lambda_i \int_X f d\mu_i + \sum_{i=1}^{\infty} \lambda_i \int_X g d\mu_i
\]
\[
= \int_X f d\mu + \int_X g d\mu
\]
where we owe the second equality to Proposition 1.2.

Let us now relax the hypothesis that \((X, \Sigma, \mu)\) is \(\sigma\)-finite. If either \(\int_X f d\mu\) or \(\int_X g d\mu\) equals \(\infty\), then we get \(\int_X (f + g) d\mu = \infty\) by the monotonicity of the Lebesgue integral, so there is nothing to prove. Assume then that both \(\int_X f d\mu\) and \(\int_X g d\mu\) are finite. For \(h \in \{f, g\}\), let us define \(A(h) := \{h = \infty\}, B_0(h) := \{h = 0\}\) and \(B_i(h) := \{i - 1 < h \leq i\}\) for each \(i \in \mathbb{N}\). Next, we put
\[
A := A(f) \cup A(g), \quad B_0 := B_0(f) \cap B_0(g) \quad \text{and} \quad B_{ij} := B_i(f) \cap B_j(g)
\]
for any nonnegative integers \(i\) and \(j\) such that \((i, j) \neq (0, 0)\). (While \(\mu(B_0) = \infty\) is possible, we certainly have \(\mu(A) = 0\) and \(\mu(B_{ij}) < \infty\) for every \(i\) and \(j\) with \((i, j) \neq (0, 0)\).) Finally, put \(S := X \setminus B_0\), and define \(\Sigma_S\) and \(\mu_S\) as in Exercise 1.16. Then, \((S, \Sigma_S, \mu_S)\) is a \(\sigma\)-finite measure space, so using what we have found in the previous paragraph, and then invoking Exercise 1.16, yield
\[
\int_S (f + g) d\mu = \int_S f d\mu + \int_S g d\mu.
\]
On the other hand, we have
\[
\int_{B_0} (f + g) d\mu = 0 = \int_{B_0} f d\mu + \int_{B_0} g d\mu.
\]
In view of Exercise 1.15, therefore, we obtain the following generalization of Proposition 1.2:

**Proposition 1.5.** Given any measure space \((X, \Sigma, \mu)\),
\[
\int_X (af + g) d\mu = a \int_X f d\mu + \int_X g d\mu
\]
for every \(a \geq 0\) and \(\Sigma\)-measurable \([0, \infty]\)-valued maps \(f\) and \(g\) on \(X\).
The convergence theorems we proved in the previous section for the expectation functional can also be generalized to the context of the Lebesgue integral by means of the same technique.

In the following set of exercises the Lebesgue integrals are taken relative to an arbitrarily fixed measure space \((X, \Sigma, \mu)\), and \((f_m)\) stands for a sequence of \([0, \infty]\)-valued \(\Sigma\)-measurable functions on \(X\).

**Exercise 1.17.** Prove: \(f_1 \leq f_2 \leq \cdots \) implies

\[
\lim \int_X f_m d\mu = \int_X \lim f_m d\mu.
\]

**Exercise 1.18.** Prove: \(\liminf \int_X f_m d\mu \leq \int_X \liminf f_m d\mu\).

**Exercise 1.19.** (The Lebesgue Convergence Theorem) Let \(f\) be a \(\Sigma\)-measurable \([0, \infty]\)-valued map on \(X\) such that \(\mu\{\omega \in X : f_m(\omega) \to f(\omega) \text{ is false}\} = 0\). Show that

\[
\int_X f d\mu = \lim \int_X f_m d\mu,
\]

if there is a \(\Sigma\)-measurable \(g : X \to [0, \infty]\) with \(\int_X g d\mu < \infty\) and \(g \geq \sup\{f, f_1, f_2, \ldots\}\).

### 1.4 Digression: Absolute Continuity

**Absolutely Continuous Distribution Functions**

Many distribution functions that arise in applications are defined through integrating certain types of nonnegative Borel measurable functions (known as “density functions”) with respect to the Lebesgue measure on an interval. The discussion of the previous section positions us well for introducing such distribution functions formally.

Let \(F\) be a distribution function. We say that \(F\) is absolutely continuous (with respect to the Lebesgue measure) if there is a Borel measurable map \(f : \mathbb{R} \to \mathbb{R}_+\) such that

\[
F(t) = \int_{(-\infty,t]} f d\ell \tag{9}
\]

for every real number \(t\). When this is the case, we say that \(F\) is induced by the density function \(f\), and refer to \(f\) as a density for \(F\). Note that, when \(f\) is a density for \(F\), the additivity of the Lebesgue integral implies

\[
\int_{\mathbb{R}} f \mathbf{1}_{(-\infty,s]} d\ell + \int_{\mathbb{R}} f \mathbf{1}_{(s,t]} d\ell = \int_{\mathbb{R}} f \mathbf{1}_{(-\infty,t]} d\ell,
\]

whence

\[
F(t) - F(s) = \int_{(s,t]} f d\ell
\]

for any real numbers \(s\) and \(t\) with \(t > s\).
Warning. Let $f$ be a density for a distribution function $F$. Then, by definition, $f$ is a Borel self-map on $\mathbb{R}$ such that $f \geq 0$. Moreover, $f$ integrates exactly to 1, because, by Example 1.4,

$$\int_{\mathbb{R}} f \, \text{d} \ell = \sum_{i \in \mathbb{Z}} \int_{(i, i+1]} f \, \text{d} \ell = \sum_{i \in \mathbb{Z}} (F(i+1) - F(i)) = F(\infty) - F(-\infty) = 1.$$ 

Conversely, any Borel measurable $f : \mathbb{R} \to \mathbb{R}_+$ with $\int_{\mathbb{R}} f \, \text{d} \ell = 1$ is a density for some distribution function. (Why?)

**Example 1.5.** For any real numbers $a$ and $b$ with $a < b$, the uniform distribution $F$ on $[a, b]$ – recall Example C.2.1 – is absolutely continuous. For instance, the map $f : \mathbb{R} \to \mathbb{R}_+$, where

$$f(t) = \begin{cases} \frac{1}{b-a}, & a \leq t \leq b \\ 0, & \text{otherwise} \end{cases}$$

is a density for that distribution function. Similarly, for any $\lambda > 0$, the exponential distribution with parameter $\lambda$ – recall Example C.2.2 – is absolutely continuous. For instance, the map $f : \mathbb{R} \to \mathbb{R}_+$, where

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

is a density for that distribution function.\footnote{Well, there is a bit of a caveat here. Equation (9) builds a connection between an absolutely continuous distribution function and its density through the Lebesgue integral on $\mathbb{R}$, and at present we don’t have any means of “computing” Lebesgue integrals efficiently. In the next chapter, I will show that one can instead use the familiar Riemann integral in this equation, provided that $f$ is “sufficiently” continuous. This will make proving the two claims I made in this example an easy exercise in calculus.}

**Example 1.6.** Some distribution functions do not have closed form descriptions, but they are rather “defined” through integrating a density function. The most important example of such a function is the so-called *normal* distribution function (with parameters $\mu$ and $\sigma$). For any given real numbers $\mu$ and $\sigma > 0$, this function $F$ is defined through equation (9), where

$$f(t) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

for any real number $t$; it is thus trivially absolutely continuous. (If $\mu = 0$ and $\sigma = 1$ here, this function is called the *standard normal* distribution function.)

A natural question at this point is if every distribution function arises from a density through the equation (9). It is easy to see that the answer is no. The
following observation shows that if a distribution function is not continuous even at a single point, then it cannot possibly possess a density.

**Proposition 1.6.** Every absolutely continuous distribution function is uniformly continuous.

**Proof.** Let \( F \) be an absolutely continuous distribution function and \( (t_m) \) a real sequence that converges to a real number \( t \). Then, an immediate application of the Lebesgue Convergence Theorem (Exercise 1.19) yields

\[
F(t) = \int_{\mathbb{R}} f1_{(-\infty,t_m]}d\ell \to \int_{\mathbb{R}} f1_{(-\infty,t]}d\ell = F(t),
\]

and we conclude that \( F \) is continuous. As every continuous distribution function is uniformly continuous (Exercise B.3.10), we are done.

The converse of Proposition 1.6 is false. Indeed, even a continuous distribution function need not possess a density.

**Example 1.7.** (The Cantor Distribution) Let \( F \) stand for the Cantor function on \([0,1]\) which was constructed in Example C.1.5. Let us extend this map to the entire \( \mathbb{R} \) by setting it equal to 0 on \((\infty,0)\) and to 1 on \((1,\infty)\). As the Cantor function is increasing and continuous, and it equals 0 at 0 and 1 at 1, the resulting map, which we also denote by \( F \), is a continuous distribution function. This distribution function is called the **Cantor distribution function**, and the Lebesgue-Stieltjes probability measure \( p_F \) induced by it on \( \mathbb{R} \) is called the **Cantor distribution**.

It is easy to see that the support of the Cantor distribution is contained in the Cantor set \( C \). Indeed, \( F \) is constant on every open interval that is deleted in the construction of the Cantor set. Adopting the notation we used in Example C.1.5, therefore, we have \( p_F(I_{m,k}) = 0 \) for any positive integer \( m \) and \( k \in [2^{m-1}] \). It follows that

\[
p_F(X \setminus C) = \sum_{m=1}^{\infty} \sum_{k \in [2^{m-1}]} p_F(I_{m,k}) = 0,
\]

that is, \( p_F(C) = 1 \). Now, to derive a contradiction, suppose there is a Borel measurable map \( f : \mathbb{R} \to \mathbb{R}_+ \) such that \( F \) satisfies (9) for every real number \( t \). Then, it must be the case that

\[
p_F(S) = \int_S f d\ell \quad \text{for every } S \in \mathcal{B}(\mathbb{R}).
\]

(Proof. Let \( G \) be the set of all \( S \in \mathcal{B}(\mathbb{R}) \) such that \( p_F(S) \) equals \( \int_S f d\ell \), and check that \( G \) is a \( \sigma \)-algebra on \( \mathbb{R} \) that contains all right semi-closed intervals.) But then \( p_F(S) = 0 \) for every \( \ell \)-null \( S \in \mathcal{B}(\mathbb{R}) \). In particular, we have \( p_F(C) = 0 \), contradicting what we found above. Conclusion: The Cantor distribution function is (uniformly) continuous, but it is not absolutely continuous.

**Exercise 1.20.** Show that every Lipschitz continuous distribution function is absolutely continuous, but not conversely.
Exercise 1.21. Let $F$ be an absolutely continuous distribution function. Show that $F(S)$ is \(\ell\)-null for every \(\ell\)-null subset $S$ of \(\mathbb{R}\).

Absolutely Continuous Measures

Let $(X, \Sigma)$ be a measurable space, and $\mu$ and $\nu$ two measures on $\Sigma$. We say that $\mu$ is absolutely continuous with respect to $\nu$—this is denoted by writing $\mu \ll \nu$—if $\mu(S) = 0$ for every $S \in \Sigma$ with $\nu(S) = 0$. For instance, the zero measure on $\Sigma$ is absolutely continuous with respect to any measure on $\Sigma$, while the counting measure on $\Sigma$ is not absolutely continuous with respect to any measure $\nu$ on $\Sigma$ such that $\nu(S) = 0$ for some nonempty $S \in \Sigma$.

The following proposition introduces a general method of obtaining absolutely continuous measures on $\Sigma$ with respect to any given measure $\nu$ on $\Sigma$.

**Proposition 1.7.** Let $(X, \Sigma)$ be a measure space and take any $f \in L^0_+(X, \Sigma)$. Then, the map $\mu : \Sigma \to \mathbb{R}$, defined by

$$
\mu(S) := \int_S f \, d\nu,
$$

is a measure on $\Sigma$ with $\mu \ll \nu$. If $\nu$ is $\sigma$-finite, so is $\mu$.

Exercise 1.22. Prove Proposition 1.7.

Exercise 1.23. Show that the Cantor distribution is not absolutely continuous with respect to the Lebesgue measure.

Exercise 1.24. Take any $\lambda > 0$, and let $p$ stand for the Lebesgue-Stieltjes probability measure induced by the exponential distribution with parameter $\lambda$ on $\mathbb{R}$. Show that $p$ is absolutely continuous with respect to the Lebesgue measure, but not conversely.

Exercise 1.25. Let $(X, \Sigma)$ be a measurable space, and let $\mu$ and $\nu$ be two finite measures on $\Sigma$ with $\mu \ll \nu$. Prove: For every $\varepsilon > 0$, there is a $\delta > 0$ such that $\mu(S) < \varepsilon$ for each $S \in \Sigma$ with $\nu(S) < \delta$.

**Warning.** This exercise shows in what sense we may think of the absolute continuity of a finite measure with respect to another such measure as a “continuity” property. However, this observation fails for infinite measures. For instance, consider the measurable space $(\mathbb{N}, 2^\mathbb{N})$ and the measures $\mu$ and $\nu$ on $2^\mathbb{N}$ with $\mu\{i\} = 2^i$ and $\nu\{i\} = 2^{-i}$ for each positive integer $i$.

Exercise 1.26. Let $(X, \Sigma)$ be a measurable space, and let $\mu$ and $\nu$ be two measures on $\Sigma$. We say that $\mu$ and $\nu$ are mutually singular, and write $\mu \perp \nu$, if there is an $S \in \Sigma$ such that $\mu(S) = 0 = \nu(X \setminus S)$. Prove:

(a) $\mu \ll \nu$ and $\mu \perp \nu$ iff $\mu = 0$;
(b) If $\mu$ and $\nu$ are finite, then $\mu \perp \nu$ iff there is no nonzero measure $\xi$ on $\Sigma$ such that $\xi \ll \mu$ and $\xi \ll \nu$;
(c) If $X$ is a metric space, then $p \perp q$ for any $p, q \in \Delta(X)$ with $\text{supp}(p) \cap \text{supp}(q) = \emptyset$.

It turns out that the method of obtaining absolutely continuous measures given in Proposition 1.7 is universal, so long as $\nu$ is $\sigma$-finite. That is, given a $\sigma$-finite measure
\( \nu \) on \( \Sigma \), any \( \sigma \)-finite measure on \( \Sigma \) that is absolutely continuous with respect to \( \nu \) arises in exactly the way prescribed in Proposition 1.7. This is the content of the following result which is one of the most important theorems of measure theory.

**The Radon-Nikodym Theorem.**\(^9\) Let \((X, \Sigma)\) be a measurable space, and \(\mu\) and \(\nu\) two \(\sigma\)-finite measures on \(\Sigma\) with \(\mu \ll \nu\). Then, there is an \(f \in L_{+}^{0}(X, \Sigma)\) such that

\[
\mu(S) = \int_{S} f \, d\nu \quad \text{for every} \ S \in \Sigma. \tag{10}
\]

**Terminology.** Any map \(f \in L_{+}^{0}(X, \Sigma)\) such that (10) holds is commonly referred to as either the **density of** \(\mu\) **with respect to** \(\nu\) or as the **Radon-Nikodym derivative of** \(\mu\) **with respect to** \(\nu\). It is quite standard to denote any such map as \(\frac{d\mu}{d\nu}\). (We will see in Exercise 1.28 that there is a sense in which this map is unique.)

**Warning.** The Radon-Nikodym Theorem is not valid for arbitrary measures. For instance, let \(X\) be a singleton set and \(\Sigma := \{\emptyset, X\}\). Let \(\mu\) and \(\nu\) be the measures on \(\Sigma\) with \(\mu(X) = 1\) and \(\nu(X) = \infty\). Then, \(\mu \ll \nu\) but there does not exist a density of \(\mu\) with respect to \(\nu\). (If there were such a density \(f\), we would have \(1 = \mu(X) = \int_{X} f \, d\nu = \infty\).)

The standard proof of the Radon-Nikodym Theorem requires a digression into the theory of **signed measures**, which we wish to avoid at this stage of our introduction. Instead, we will refrain from using this result in this text (with one major exception in the next chapter), and provide a proof for it only in Chapter L as an application of our subsequent work on probability theory.

In the following set of exercises \((X, \Sigma)\) stands for an arbitrary measurable space.

**Exercise 1.27.**\(^1\) Let \(\nu\) stand for the counting measure on \(B[0,1]\). Show that \(\ell \ll \nu\), but there does not exist a density of \(\ell\) with respect to \(\nu\). (This is another example that shows that the Radon-Nikodym Theorem fails for non-\(\sigma\)-finite measures.)

**Exercise 1.28.** Let \(\mu\) and \(\nu\) be two \(\sigma\)-finite measures on \(\Sigma\). Prove: If \(f\) and \(g\) are both densities of \(\mu\) with respect to \(\nu\), then \(\nu\{f \neq g\} = 0\).

**Exercise 1.29.** Let \(\mu_1, \mu_2\) and \(\nu\) be finite measures on \(\Sigma\) such that \(\mu_1 \ll \nu\) and \(\mu_2 \ll \nu\). Then \(\mu_1 + \mu_2 \ll \nu\), so the Radon-Nikodym derivatives of \(\mu_1 + \mu_2\), \(\mu_1\) and \(\mu_2\) with respect to \(\nu\) exist. Prove that \(\nu\left\{\frac{d(\mu_1 + \mu_2)}{d\nu} \neq \frac{d\mu_1}{d\nu} + \frac{d\mu_2}{d\nu}\right\} = 0\).

**Exercise 1.30.** Let \(\mu, \mu_0\) and \(\nu\) be finite measures on \(\Sigma\) such that \(\mu \ll \mu_0\) and \(\mu_0 \ll \nu\). Prove that \(\nu\left\{\frac{d\mu}{d\nu} \neq \frac{d\mu}{d\mu_0} \frac{d\mu_0}{d\nu}\right\} = 0\).

---

\(^9\)This theorem was proved by Johann Radon in 1913 in the special case where \((X, \Sigma)\) is \((\mathbb{R}^n, B(\mathbb{R}^n))\). It was obtained in its full glory by Otto Nikodym in 1930.
Exercise 1.31. Let $\mu$ and $\nu$ be finite measures on $\Sigma$, and $f \in L^0_+(X, \Sigma)$. Prove that

$$\int_S f \, d\mu = \int_S f \frac{d\mu}{d\nu} \, d\nu \quad \text{for every } S \in \Sigma.$$  

Exercise 1.32. (The Lebesgue Decomposition Theorem) Let $\mu$ and $\nu$ be finite measures on $\Sigma$. In this exercise, we prove that there are measures $\mu_1$ and $\mu_2$ on $\Sigma$ such that $\mu_1 \ll \nu$, $\mu_2 \perp \nu$ and $\mu = \mu_1 + \mu_2$.

(a) Use the Radon-Nikodym Theorem to show that there exist a density of $\mu$ with respect to $\mu + \nu$, say $g$, and a density of $\nu$ with respect to $\mu + \nu$, say $h$.

(b) Put $Y := \{ h \neq 0 \}$, and define $f := \frac{g}{h} \mathbb{1}_Y$, where $\mathbb{1}_Y$ is the indicator function of $Y$ on $X$. Next, define the real maps $\mu_1$ and $\mu_2$ on $\Sigma$ by

$$\mu_1(S) := \int_S f \, d\nu \quad \text{and} \quad \mu_2(S) := \mu(S \cap (X \setminus Y)).$$

Check that $\mu_1 \ll \nu$ and $\mu_2 \perp \nu$.

(c) Use Exercise 1.31 to verify that $\mu = \mu_1 + \mu_2$.

In passing, we note that the notion of absolute continuity we have defined for measures is more general than that we have defined for distribution functions, provided that we apply the latter to the induced Lebesgue-Stieltjes probability measures. The following result clarifies this situation.

**Proposition 1.8.** Let $F$ be a distribution function and $p_F$ the Lebesgue-Stieltjes probability measure on $\mathbb{R}$ induced by $F$. Then, $F$ is absolutely continuous if, and only if, $p_F$ is absolutely continuous with respect to $\ell$.

**Proof.** Suppose that $F$ is absolutely continuous. Then, there is a Borel measurable map $f : \mathbb{R} \to \mathbb{R}_+$ such that $F(t) = \int_{(-\infty,t]} f \, d\ell$ for every real number $t$. Let us define $\mu : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ by $\mu(S) := \int_S f \, d\ell$. By Exercise 1.17,

$$\mu(\mathbb{R}) = \lim_{m \to \infty} \int_{(-\infty,m]} f \, d\ell = F(\infty) = 1,$$

so, in view of Proposition 1.7, $\mu$ is a Borel probability measure on $\mathbb{R}$ such that $\mu \ll \ell$. But, as $p_F(-\infty, t] = \mu(-\infty, t]$ for each $t \in \mathbb{R}$, Proposition B.4.1 entails $p_F = \mu$, so $p_F \ll \ell$, as we desired. Conversely, suppose $p_F \ll \ell$. Then, by the Radon-Nikodym Theorem, there is a density of $p_F$ with respect to $\ell$, say, $f$. Hence,

$$F(t) = p_F(-\infty, t] = \int_{(-\infty,t]} f \, d\ell$$

for every real number $t$, which means that $F$ is absolutely continuous. \hfill \blacksquare

23
2 The Expectation Functional II

2.1 Expectation of Arbitrary Random Variables

Expectation by Additivity w.r.t. Positive and Negative Parts

We now wish to extend the expectation functional to the collection of all random variables on a given probability space. And, we would like to do this in a way that preserves the linearity property of this functional. Let \( x \) be an \( \mathbb{R} \)-valued random variable on a probability space \( (X, \Sigma, \mathbf{p}) \); our problem is to find a definition for \( \mathbf{E}(x) \). Recall that \( x^+ := \max\{x, 0\} \) and \( x^- := \max\{-x, 0\} \) are \([0, \infty]\)-valued random variables on \((X, \Sigma, \mathbf{p})\) with \( x = x^+ - x^- \) (Section C.1.2). Then, whatever is its definition, \( \mathbf{E}(x) \) must satisfy \( \mathbf{E}(x) = \mathbf{E}(x^+) - \mathbf{E}(x^-) \). Aha! As \( x^+ \) and \( x^- \) are \([0, \infty]\)-valued, the right-hand side of this expression is already defined. Therefore, the linearity property that we desire on \( \mathbf{E} \) forces us to define the expectation of \( x \) as the (extended real) number

\[
\mathbf{E}(x) := \mathbf{E}(x^+) - \mathbf{E}(x^-),
\]

provided that \( \mathbf{E}(x^+) \) and \( \mathbf{E}(x^-) \) are not both infinite. (Clearly, this definition generalizes the one we gave for the expectation of \([0, \infty]\)-valued random variables.) If \( \mathbf{E}(x^+) = \infty = \mathbf{E}(x^-) \), we say that the expectation of \( x \) does not exist. (In real analysis, \( \mathbf{E}(x) \) is called the Lebesgue integral of \( x \) (with respect to \( \mathbf{p} \)), and is denoted by \( \int_X x \, d\mathbf{p} \) – see Section 2.2.) In turn, we define the variance of \( x \) as the (extended real) number

\[
\mathbf{V}(x) := \mathbf{E}((x - \mathbf{E}(x))^2),
\]

provided that \( \mathbf{E}(x) \) is finite. (Here, of course, \( x - \mathbf{E}(x) \) is the map \( \omega \mapsto x(\omega) - \mathbf{E}(x) \) on \( X \).)

Warning. It is not at all weird for a random variable not to have an expectation. For instance, consider the probability space \((\mathbb{N}, 2^\mathbb{N}, \mathbf{p})\) where \( \mathbf{p}(\{i\}) := 2^{-i} \) for each \( i \in \mathbb{N} \), and define the random variable \( x \) on this space by setting \( x(i) := 2^i \) if \( i \) is odd, and \( x(i) := -2^{-i} \) if \( i \) is even. It is readily verified that \( \mathbf{E}(x^+) = \infty = \mathbf{E}(x^-), \) so \( \mathbf{E}(x) \) does not exist. \(^{10}\)

Integrability

By definition, \( \mathbf{E}(x) \) exists iff \( \min\{\mathbf{E}(x^+), \mathbf{E}(x^-)\} < \infty \). As \( |x| = x^+ + x^- \), Proposition 1.2 says that

\[
\mathbf{E}(|x|) = \mathbf{E}(x^+) + \mathbf{E}(x^-),
\]

so we see that \( \mathbf{E}(|x|) \) is finite iff \( \max\{\mathbf{E}(x^+), \mathbf{E}(x^-)\} < \infty \). When this is the case, that is, \( \mathbf{E}(|x|) < \infty \), we say that \( x \) is integrable (with respect to \( \mathbf{p} \)). In other words, \( x \) is integrable iff \( \mathbf{E}(x) \) exists and it is finite.

\(^{10}\)There are less artificial examples of random variables without expectations. Indeed, even a random variable with an absolutely continuous distribution function may fail to have an expectation. I will give an example of this sort in the next chapter.
Insight. For an \( \mathbb{R} \)-valued random variable \( x \), \( \mathbb{E}(x) \) exists iff either \( \mathbb{E}(x^+) \) or \( \mathbb{E}(x^-) \) is finite. On the other hand, \( x \) is integrable iff both \( \mathbb{E}(x^+) \) and \( \mathbb{E}(x^-) \) are finite.

The set of all integrable random variables on a probability space \((X, \Sigma, \mathbf{p})\) is denoted by \( \mathcal{L}^1(X, \Sigma, \mathbf{p}) \), that is,

\[
\mathcal{L}^1(X, \Sigma, \mathbf{p}) := \left\{ x \in \mathcal{L}^0(X, \Sigma) : \int_X |x| \, d\mathbf{p} < \infty \right\}.
\]

It is plain that \( ax \in \mathcal{L}^1(X, \Sigma, \mathbf{p}) \) for any real number \( a \) and \( x \in \mathcal{L}^1(X, \Sigma, \mathbf{p}) \). Besides, if \( x \) and \( y \) are integrable random variables on \((X, \Sigma, \mathbf{p})\), using the Triangle Inequality (for real numbers), and Propositions 1.1 and 1.2, we find

\[
\mathbb{E}(|x + y|) \leq \mathbb{E}(|x| + |y|) = \mathbb{E}(|x|) + \mathbb{E}(|y|) < \infty,
\]

that is, \( x + y \in \mathcal{L}^1(X, \Sigma, \mathbf{p}) \). Conclusion: \( \mathcal{L}^1(X, \Sigma, \mathbf{p}) \) is a linear subspace of \( \mathcal{L}^0(X, \Sigma) \).

Summary

We finally have the full definition of the expectation functional at our disposal. Before proceeding any further, let us briefly review how we arrived at it. We first defined \( \mathbb{E} \) for simple random variables on a given probability space \((X, \Sigma, \mathbf{p})\) in an intuitive way (using the weighted average idea), and then extended it to the class of all \([0, \infty]\)-valued random variables on \((X, \Sigma, \mathbf{p})\) by means of approximation from below. We managed to preserve the linearity of \( \mathbb{E} \) during this extension. This was not a trivial thing to do, and it required the combined power of the Monotone Convergence Theorem 1 and the fact that any \([0, \infty]\)-valued random variable on \((X, \Sigma, \mathbf{p})\) can be approximated (in fact, uniformly) from below by an increasing sequence of simple random variables on \((X, \Sigma, \mathbf{p})\). Finally, we have extended \( \mathbb{E} \) further to the collection of all \( \mathbb{R} \)-valued random variables by means of applying it to the positive and negative parts of random variables in an additive way. It should not be surprising that linearity of \( \mathbb{E} \) is preserved in this extension as well (but this still needs to be proved).

Example 2.1. Let \((X, 2^X, \mathbf{p})\) be a probability space with \( X \) being a countably infinite set. We wish to find an expression for the expectation of an arbitrary integrable random variable \( x \) on \((X, 2^X, \mathbf{p})\). Integrability of \( x \) means that both \( \mathbb{E}(x^+) \) and \( \mathbb{E}(x^-) \) are finite. Therefore, by Example 1.3 and Corollary 1.9 of Appendix 1,

\[
\mathbb{E}(x^+) - \mathbb{E}(x^-) = \sum_{\omega \in X} x^+(\omega)\mathbf{p}\{\omega\} - \sum_{\omega \in X} x^-(\omega)\mathbf{p}\{\omega\} = \sum_{\omega \in X} (x^+ - x^-)(\omega)\mathbf{p}\{\omega\},
\]

that is,

\[
\mathbb{E}(x) = \sum_{\omega \in X} x(\omega)\mathbf{p}\{\omega\}
\]
for every $x \in \mathcal{L}^1(X, 2^X, p)$.

\section*{Properties of $\mathbb{E}$}

Let $(X, \Sigma, p)$ be a probability space, and $x$ an $\mathbb{F}$-valued random variable on $(X, \Sigma, p)$. Assume that $\mathbb{E}(x)$ exists. Then $|\mathbb{E}(x)|$ is an element of $[0, \infty]$. How does this number compare with $\mathbb{E}(|x|)$? Well, since $\mathbb{E}$ is a type of integral, it is natural to expect that $|\mathbb{E}(x)| \leq \mathbb{E}(|x|)$. There is no surprise in this regard. We have

$$|\mathbb{E}(x)| = |\mathbb{E}(x^+) - \mathbb{E}(x^-)| \leq |\mathbb{E}(x^+)| + |\mathbb{E}(x^-)| = \mathbb{E}(x^+) + \mathbb{E}(x^-)$$

while, by Proposition 1.2, $\mathbb{E}(x^+) + \mathbb{E}(x^-) = \mathbb{E}(x^+ + x^-) = \mathbb{E}(|x|)$. We conclude:

**Proposition 2.1.** For any $\mathbb{F}$-valued random variable $x$ on a probability space $(X, \Sigma, p)$ such that $\mathbb{E}(x)$ exists, we have $|\mathbb{E}(x)| \leq \mathbb{E}(|x|)$.

Let us now turn to the linearity of $\mathbb{E}$. The idea is to deduce this property from the linear behavior of $\mathbb{E}$ over $[0, \infty]$-valued random variables. Doing this is not difficult, but it is a tad bit tedious.

**Proposition 2.2.** Let $x$ and $y$ be $\mathbb{F}$-valued random variables on a probability space $(X, \Sigma, p)$. For any real number $a$, we have $\mathbb{E}(ax) = a\mathbb{E}(x)$, provided that $\mathbb{E}(x)$ exists. Furthermore, $\mathbb{E}(x + y) = \mathbb{E}(x) + \mathbb{E}(y)$, provided that $x + y \in \mathbb{F}^X$, $\mathbb{E}(x)$ exists, and $y$ is integrable.

**Proof.** For any real number $a$, it is easily verified that

$$(ax)^+ = \begin{cases} 
ax^+, & \text{if } a \geq 0 \\
(-a)x^-, & \text{if } a < 0
\end{cases}$$

and

$$(ax)^- = \begin{cases} 
ax^-, & \text{if } a \geq 0 \\
(-a)x^+, & \text{if } a < 0
\end{cases}$$

and hence, by Proposition 1.2,

$$\mathbb{E}((ax)^+) = \begin{cases} 
\mathbb{E}(x^+), & \text{if } a \geq 0 \\
(-a)\mathbb{E}(x^-), & \text{if } a < 0
\end{cases}$$

and

$$\mathbb{E}((ax)^-) = \begin{cases} 
\mathbb{E}(x^-), & \text{if } a \geq 0 \\
(-a)\mathbb{E}(x^+), & \text{if } a < 0
\end{cases}$$

Since either $\mathbb{E}(x^+)$ or $\mathbb{E}(x^-)$ is finite, therefore, we find that $\mathbb{E}(ax)$ exists, and adding $\mathbb{E}((ax)^+)$ and $-\mathbb{E}((ax)^-)$ yields $a\mathbb{E}(x)$.

To prove our second assertion, let us first show that $\mathbb{E}(x + y)$ exists. We are given that $x + y$ is a well-defined $\mathbb{F}$-valued random variable on $(X, \Sigma, p)$. Furthermore, $(x + y)^+ \leq x^+ + y^+$ and $(x + y)^- \leq x^- + y^-$, which is readily verified by using the definitions of the positive and negative parts of an $\mathbb{F}$-valued function. It then follows from Propositions 1.1 and 1.2 that

$$\mathbb{E}((x + y)^+) \leq \mathbb{E}(x^+) + \mathbb{E}(y^+) \quad \text{and} \quad \mathbb{E}((x + y)^-) \leq \mathbb{E}(x^-) + \mathbb{E}(y^-). \quad (11)$$
But we know that either $E(x^+)$ or $E(x^-)$ is finite and that both $E(y^+)$ and $E(y^-)$ are finite. Therefore, either $E((x+y)^+)$ and $E((x+y)^-)$ is finite, that is, $E(x+y)$ exists.

We now make note of the following identity:\footnote{When both $x$ and $y$ are real-valued, this identity comes from decomposing $x+y$ in two ways:}

$$(x+y)^+ + x^- + y^- = (x+y)^- + x^+ + y^+. \tag{12}$$

Then, using the additivity of $E$ over $[0, \infty]$-valued random variables, we get

$$E((x+y)^+) + E(x^-) + E(y^-) = E((x+y)^-) + E(x^+) + E(y^+). \tag{13}$$

There are three cases to look at. First, consider the possibility $E(x^+) < \infty$ and $E(x^-) = \infty$. In this case, since $E(y^+)$ is finite, (13) implies $E((x+y)^-) = \infty$. Thus, given that $E(x+y)$ exists, we have $E(x+y) = -\infty = E(x) + E(y)$.

If, on the other hand, $E(x^+) = \infty$ and $E(x^-) < \infty$, then the analogous reasoning shows that $E(x+y)$ and $E(x^+) + E(y)$ both equal $\infty$. Finally, if both $E(x^+)$ and $E(x^-)$ are finite, then (11) implies that both $E((x+y)^+)$ and $E((x+y)^-)$ are finite. Thus, in this case all terms in the equation (13) are finite, so we can shuffle things back to get

$$E((x+y)^+) - E((x+y)^-) = E(x^+) - E(x^-) + E(y^+) - E(y^-),$$

that is, $E(x+y) = E(x) + E(y)$, as we sought.

\textbf{Insight.} The set of all integrable random variables on a probability space $(X, \Sigma, \mathbf{p})$, that is, $\mathcal{L}^1(X, \Sigma, \mathbf{p})$, is a linear space (relative to the usual scalar multiplication and addition operations). The map $E$ acts as a linear functional on this linear space.\footnote{$E$ is not a linear functional on $\mathcal{L}^0(X, \Sigma)$, because we cannot write $E(x-y) = E(x) - E(y)$ for $x,y \in \mathcal{L}^0(X, \Sigma)$ if the expectations of both $x$ and $y$ are $\infty$. When we restrict the domain of $E$ to $\mathcal{L}^1(X, \Sigma, \mathbf{p})$, however, this difficulty disappears. (I will talk more about this issue in Section 5.)}

In the following set of exercises the expectations are taken relative to an arbitrarily fixed probability space $(X, \Sigma, \mathbf{p})$.

\textbf{Exercise 2.1.} Let $a$ and $b$ be two real numbers with $a \leq b$, and $x$ an $[a,b]$-valued random variable on $(X, \Sigma, \mathbf{p})$. Show that $E(x)$ exists, and $a \leq E(x) \leq b$. 

[27]
Exercise 2.2.\textsuperscript{H} (Monotonicity of $\mathbb{E}$) Let $x$ and $y$ be two integrable $\mathbb{R}$-valued random variables on $(X, \Sigma, p)$. Show that

$$x \geq y \quad \text{implies} \quad \mathbb{E}(x) \geq \mathbb{E}(y),$$

and

$$x \geq y \quad \text{and} \quad \mathbb{E}(x) = \mathbb{E}(y) \quad \text{imply} \quad x = y.$$

Another Proof for Proposition 2.1. Suppose $\mathbb{E}(x)$ exists. If $\mathbb{E}(x) \in \{-\infty, \infty\}$, then $\mathbb{E}(|x|) = \infty$, so our claim holds. On the other hand, if $x$ is integrable, we may use the monotonicity of $\mathbb{E}$ to find $\mathbb{E}(x) \leq \mathbb{E}(|x|)$ and $-\mathbb{E}(x) = \mathbb{E}(-x) \leq \mathbb{E}(|x|)$, and it follows again that $|\mathbb{E}(x)| \leq \mathbb{E}(|x|)$.

Exercise 2.3.\textsuperscript{H} For any $x, y \in L^1(X, \Sigma, p)$, show that if $\mathbb{E}(x1_S) \geq \mathbb{E}(y1_S)$ for every $S \in \Sigma$, then $x \geq_{\text{a.s.}} y$.

Exercise 2.4.\textsuperscript{H} Let $\{S_i \in \Sigma : i = 1, 2, \ldots\}$ be a partition of $X$. Show that

$$\int_X xdp = \sum_{i=1}^{\infty} \int_{S_i} xdp \quad \text{for every} \quad x \in L^1(X, \Sigma, p).$$

Exercise 2.5.\textsuperscript{H} Let $n$ be a positive integer, and let $L^{1,n}(X, \Sigma, p)$ stand for the set of all $\mathbb{R}^n$-valued random variables on $(X, \Sigma, p)$ such that $\mathbb{E}(|f \circ x|) < \infty$ for every linear functional $f$ on $\mathbb{R}^n$. Show that for any $x$ in $L^{1,n}(X, \Sigma, p)$, there is a unique $n$-vector $v_x$ such that $f(v_x) = \mathbb{E}(f \circ x)$ for every linear functional $f$ on $\mathbb{R}^n$, and express $v_x$ by using the expectation functional $\mathbb{E}$.

Monotonicity of $\mathbb{E}$

We can do better than the monotonicity result reported in Exercise 2.2. Indeed, in view of Proposition 1.1, a natural conjecture is that $x \geq_{\text{a.s.}} y$ is enough to ensure $\mathbb{E}(x) \geq \mathbb{E}(y)$, at least when $x$ and $y$ are integrable. Not only that this conjecture is correct, but we do not even need an integrability requirement for it.

\textbf{Proposition 2.3.} Let $x$ and $y$ be two $\mathbb{R}$-valued random variables (on a given probability space). If both $\mathbb{E}(x)$ and $\mathbb{E}(y)$ exist, then

$$x \geq_{\text{a.s.}} y \quad \text{implies} \quad \mathbb{E}(x) \geq \mathbb{E}(y). \quad (14)$$

\textbf{Proof.} Suppose $x \geq_{\text{a.s.}} y$, and note that this entails $x^+ \geq_{\text{a.s.}} y^+$ and $y^- \geq_{\text{a.s.}} x^-$. It thus follows from Proposition 1.1 that $\mathbb{E}(x^+) \geq \mathbb{E}(y^+)$ and $\mathbb{E}(y^-) \geq \mathbb{E}(x^-)$. So, if both $\mathbb{E}(x)$ and $\mathbb{E}(y)$ exist, we have $\mathbb{E}(x) \geq \mathbb{E}(y)$. \hfill \blacksquare

\textbf{Insight.} The linear space $L^1(X, \Sigma, p)$ becomes a preordered linear space when endowed with the preorder $\geq_{\text{a.s.}}$. The map $\mathbb{E}$ acts as an increasing linear functional on this space.\textsuperscript{13}

\textsuperscript{13}Quiz. Show that $\mathbb{E}$ is actually strictly increasing on this space.
A moment’s thought shows that if two \( \mathbb{R} \)-valued random variables \( x \) and \( y \) are almost surely equal, then \( x^+ =_{a.s.} y^+ \) and \( x^- =_{a.s.} y^- \), and hence, \( \mathbb{E}(x) \) exists iff \( \mathbb{E}(y) \) exists. An immediate consequence of Proposition 2.3 is thus:

**Corollary 2.4.** Let \( x \) and \( y \) be two \( \mathbb{R} \)-valued random variables (on a given probability space). If \( x =_{a.s.} y \) and \( \mathbb{E}(x) \) exists, then \( \mathbb{E}(x) = \mathbb{E}(y) \).

The existence of both \( \mathbb{E}(x) \) and \( \mathbb{E}(y) \) is necessary for (14) to make sense. Formally, then, this is a minimal condition that is needed for Proposition 2.3 to work. But it is sometimes easier to deduce the validity of this condition by checking for stronger hypotheses. In particular, the following fact is quite useful.

**Corollary 2.5.** Let \( x \) and \( y \) be two \( \mathbb{R} \)-valued random variables (on a given probability space) such that \( y \geq_{a.s.} x \). If \( y \) is integrable, or \( \mathbb{E}(x) > -\infty \), then, both \( \mathbb{E}(x) \) and \( \mathbb{E}(y) \) exist, and we have \( \mathbb{E}(y) \geq \mathbb{E}(x) \).

**Proof.** As \( y \geq_{a.s.} x \) implies \( y^+ \geq_{a.s.} x^+ \) and \( x^- \geq_{a.s.} y^- \), we have \( \mathbb{E}(y^+) \geq \mathbb{E}(x^+) \) and \( \mathbb{E}(x^-) \geq \mathbb{E}(y^-) \) by Proposition 1.1. It follows that \( \mathbb{E}(x^+) \) is finite if \( y \) is integrable, and \( \mathbb{E}(y^-) \) is finite if \( \mathbb{E}(x) > -\infty \). Therefore, both \( \mathbb{E}(x) \) and \( \mathbb{E}(y) \) exist under either of our hypotheses, and applying Proposition 2.3 completes the proof. \( \blacksquare \)

**Monotone Convergence Theorem, Again**

The next order of business is to extend the Monotone Convergence Theorem 1 to the context of all \( \mathbb{R} \)-valued random variables.

**The Monotone Convergence Theorem 2.** Let \( x, x_1, x_2, \ldots \) be \( \mathbb{R} \)-valued random variables on a probability space \( (X, \Sigma, \mathbb{P}) \). Then,

\[
\mathbb{E}(x_1) > -\infty \text{ and } x_m \uparrow_{a.s.} x \text{ imply } \mathbb{E}(x_m) \uparrow \mathbb{E}(x).
\]

**Proof.** Assume \( \mathbb{E}(x_1) > -\infty \) and \( x_m \uparrow_{a.s.} x \). Then, by Corollary 2.5, \( \mathbb{E}(x) \) and each \( \mathbb{E}(x_m) \) exist, and we have \( \mathbb{E}(x) \geq \cdots \geq \mathbb{E}(x_2) \geq \mathbb{E}(x_1) > -\infty \). Thus, if \( \mathbb{E}(x_m) = \infty \) for some \( m \), our assertion is trivially true. Let us assume then that \( \mathbb{E}(x_m) < \infty \) for each \( m = 1, 2, \ldots \), which, in particular, implies that \( x_1 \) is integrable.

Let us suppose for a moment that \( x_1 \) is real-valued and \( \cdots \geq x_2 \geq x_1 \). Then, \( x_m - x_1 \) is a \( [0, \infty] \)-valued random variable on \( (X, \Sigma, \mathbb{P}) \) for each \( m \), and we have \( x_m - x_1 \uparrow_{a.s.} x - x_1 \). Therefore, by the linearity of the expectation functional (Proposition 2.2) and the Monotone Convergence Theorem 1,

\[
\mathbb{E}(x) - \mathbb{E}(x_1) = \mathbb{E}(x - x_1) = \lim \mathbb{E}(x_m - x_1) = \lim \mathbb{E}(x_m) - \mathbb{E}(x_1).
\]

As \( \mathbb{E}(x_1) \) is finite, we thus get \( \mathbb{E}(x) = \lim \mathbb{E}(x_m) \), as desired.
We have now proved our result in the case where $x_1$ is real-valued and \( \cdots \geq x_2 \geq x_1 \). We now drop these assumptions by using our usual “trick.” Put \( S := \{ x_1 \in \mathbb{R} \text{ and } \cdots \geq x_2 \geq x_1 \} \). Not only does this set belong to \( \Sigma \), but we have \( p(S) = 1 \) because \( x_1 \) is integrable and \( \cdots \geq a.s. \ x_2 \geq a.s. \ x_1 \). Next, we define \( x^* := x1_S \) and \( x_m^* := x_m1_S \) for each \( m \). Then, \( x^*_m \) is a (real-valued) random variable, while \( x^* \) and \( x^*_2, x^*_3, \ldots \) are \( \mathbb{R} \)-valued random variables, on \( (X, \Sigma, p) \). Furthermore, \( x^* = a.s. \ x \) and \( x_m^* = a.s. \ x_m \) for each \( m \), so, by Corollary 2.4, \( E(x^*) = E(x) \) and \( E(x_m^*) = E(x_m) \) for each \( m \). In particular, \( E(x_1^*) = E(x_1) > -\infty \), so we may apply what we have established above to \( x^*, x_1^*, x_2^*, \ldots \) to conclude that

\[
E(x) = E(x^*) = \lim E(x_m^*) = \lim E(x_m).
\]

This completes our proof.

**Warning.** The condition \( E(x_1) > -\infty \) cannot be omitted in the statement of the above theorem (although it can obviously be relaxed to \( E(x_k) > -\infty \) for some \( k \)). To see this, consider the probability space \( (N, 2^N, p) \) where \( p(i) := 2^{-i} \) for every \( i \in N \). For each positive integer \( m \), define \( x_m \in \mathcal{L}^0(N, 2^N) \) by \( x_m(i) := -\frac{1}{2^i} \). By using the observation noted in Example 1.3 (and the first part of Proposition 2.2), we find

\[
E(x_m) = -\int_N -x_mdp = \sum_{i=1}^\infty x_m(i)p(i) = -\sum_{i=1}^\infty \frac{1}{2^i} = -\infty
\]

for each \( m \), even though \( x_m \uparrow 0 \). Thus, in this instance, we have \( \lim E(x_m) = -\infty < 0 = E(\lim x_m) \).

**Dominated Convergence Theorem**

There are also some convergence theorems for the expectation functional that are applicable to non-monotonic sequences of random variables. In particular, as we shall see shortly, we can indeed interchange the operations of “applying \( E \)” and “taking the pointwise limit” for any sequence of random variables \( (x_m) \) whose absolute values are uniformly bounded (in the sense that \( \sup\{||x_1||, ||x_2||, \ldots \} < \infty \)). We can actually do better than this; all we really need is that \( y \geq a.s. \sup\{||x_1||, ||x_2||, \ldots \} \) for some integrable random variable. The following version of Fatou’s Lemma for \( \mathbb{R} \)-valued random variables play a key role in this story.

**Lemma 2.6.** (Fatou) Let \( (X, \Sigma, p) \) be a probability space, and take any \( \mathbb{R} \)-valued random variables \( y, x_1, x_2, \ldots \) on \( (X, \Sigma, p) \). Assume that \( y \) is integrable and \( |x_m| \leq a.s. \ y \) for each \( m \). Then

\[
E(\lim inf x_m) \leq \lim inf E(x_m)
\]

and

\[
\lim sup E(x_m) \leq E(\lim sup x_m).\tag{15}
\]

\( ^{14} \)That the expectations of \( \lim inf x_m \) and \( \lim sup x_m \) exist is part of the assertion.
Proof. We follow exactly the same strategy we used earlier to deduce Fatou’s Lemma from the Monotone Convergence Theorem 1; we just have to make sure that the requirements of Proposition 2.3 and the Monotone Convergence Theorem 2 are met. So, we begin by defining $y_m := \inf \{ x_m, x_{m+1}, \ldots \}$, and noting that $y_m \in C^{0}(X, \Sigma)$ for each positive integer $m$ (Exercise C.1.8). Since $y \geq a.s. |x_k|$ holds for each $k$, Boole’s Inequality ensures that $p\{ x_k > y \text{ for some } k \} = 0$, and it follows that

$$y \geq \sup_{a.s.} \{ |x_1|, |x_2|, \ldots \} \geq |y_m|$$

for any $m$. As $y$ is integrable, therefore, Corollary 2.5 ensures that each $x_m$ and $y_m$ are integrable. Given that $y_m \uparrow \liminf x_m$ we may thus apply the Monotone Convergence Theorem 2 to find

$$\int_X \liminf x_m \, dp = \lim \int_X y_m \, dp \leq \liminf \left( \int_X x_m \, dp \right)$$

where the inequality follows from the fact that $y_m \leq x_m$ for each $m$, and Proposition 2.3. To prove our second claim, we use the first claim to find $E(\liminf(-x_m)) \leq \liminf E(-x_m)$, and multiply both sides of this inequality by $-1$.\footnote{Reminder. $\limsup a_m = -\liminf(-a_m)$ for any sequence $(a_m)$ of extended real numbers.}

The following is another convergence theorem of major importance. This result provides a useful alternative to the Monotone Convergence Theorem 2, because it works for certain non-monotonic sequences of random variables. All it requires is that the sequence under consideration be “dominated” by an integrable random variable in a suitable sense.

The Dominated Convergence Theorem. (Lebesgue) Let $x, x_1, x_2, \ldots$ be $\mathbb{R}$-valued random variables on a probability space $(X, \Sigma, \mathbb{P})$. If there exists an integrable $\mathbb{R}$-valued random variable $y$ on $(X, \Sigma, \mathbb{P})$ such that $|x_m| \leq a.s. y$ for each $m$, then

$$x_m \to x \text{ a.s. } \implies E(x_m) \to E(x).$$

Proof. Let us begin by noting that the integrability of $y$, and the hypothesis $|x_m| \leq a.s. y$, guarantee that $x_m$ is integrable for each $m$. If $x_m \to_{a.s.} x$, we have $|x| \leq a.s. y$, so $x$ is integrable as well. (Yes?) Assume now that $x_m \to x$. Then $\liminf x_m = x = \limsup x_m$, so, by Lemma 2.6,

$$E(x) \leq \liminf E(x_m) \leq \limsup E(x_m) \leq E(x),$$

that is, $\lim E(x_m) = E(x)$. Now consider the general case in which $x_m \to_{a.s.} x$. Define $S := \{ x_m \to x \}$, and note that $x_{1S} = a.s. x$ and $x_{m1S} = a.s. x_m$ for each $m$. By Corollary 2.4, therefore, $E(x_{1S}) = E(x)$ and $E(x_{m1S}) = E(x_m)$ for each $m$. Furthermore, we
obviously have \( x_m 1_S \to x 1_S \), so, by what we have already proved, \( \mathbb{E}(x_m 1_S) \to \mathbb{E}(x 1_S) \), and hence \( \mathbb{E}(x_m) \to \mathbb{E}(x) \), as we sought.

**Insight.** Suppose you wish to verify that the limit of the expectations of a given sequence of random variables equals the expectation of an almost sure limit random variable. Loosely speaking, there are two standard methods of doing this. First, check if your sequence is almost surely increasing (or decreasing), and use a suitable version of the *monotone convergence theorem*. If this does not work, check if the absolute values of your random variables are bounded above by an integrable random variable almost surely. If this is the case, then use the *dominated convergence theorem*.\(^{16}\)

**Example 2.2.** For any Borel probability measure \( \mathbf{p} \) on \( \mathbb{R} \), we have \( \mathbf{p}(m, m + 1] \to 0 \).

An indirect proof of this would go something like this: If \( \mathbf{p}(m, m + 1] \to 0 \) is false, then there exists a strictly increasing sequence \( (m_k) \) of positive integers and an \( \varepsilon > 0 \) such that \( \mathbf{p}(m_k, m_k + 1] \geq \varepsilon \) for each \( k \), and hence \( 1 = \mathbf{p}(\mathbb{R}) \geq \mathbf{p}(m_1, m_1 + 1] + \mathbf{p}(m_2, m_2 + 1] + \cdots = \infty \), a contradiction. Alternatively, we can provide a direct proof by means of a dominated convergence argument. Indeed, as \( 1_{(m,m+1]} \leq 1 \) for each \( m \), and \( 1_{(m,m+1]} \to 0 \), we have

\[
\mathbf{p}(m, m + 1] = \int_{\mathbb{R}} 1_{(m,m+1]} d\mathbf{p} \to 0
\]

by the Dominated Convergence Theorem.

**Example 2.3.** Let \( (X, \Sigma, \mathbf{p}) \) be a probability space, \( I \) an interval, and \( x \) a random variable on \( X \) such that

(i) \( x(\cdot, t) \in L^1(X, \Sigma, \mathbf{p}) \) for each \( t \in I \); and

(ii) \( x(\omega, \cdot) \in C(I) \) for each \( \omega \in X \).

Consider the real map \( h \) on \( I \) defined by

\[
h(t) := \int_X x(\cdot, t) d\mathbf{p}.
\]

Question: Is \( h \) continuous? (The hypothesis (i) makes this problem well-defined, and (ii) gives it a fighting chance.) The answer is in general no, but we can use a dominated convergence argument to provide a useful sufficient condition to get a positive answer. Indeed, suppose:

(iii) There is a \( y \in L^1(X, \Sigma, \mathbf{p}) \) with \( |x(\cdot, t)| \leq_{a.s.} y \) for each \( t \in I \).

\(^{16}\)There are also hybrid versions of these theorems. For instance, the theorem I gave in Exercise 1.7 is of this form. That result is covered neither by the Monotone Convergence Theorem 1 (because \( (x_m) \) is not increasing there) nor by the Dominated Convergence Theorem 1 (because the dominating random variable there need not be integrable).
For any $t \in I$, and any sequence $(t_m)$ in $I$ with $t_m \to t$, we have $x(\cdot, t_m) \to x(\cdot, t)$ by (ii). Therefore, given (iii), we may apply the Dominated Convergence Theorem to find $h(t_m) \to h(t)$. We thus conclude that $h$ is continuous under the additional hypothesis (iii). □

**Young’s Theorem**

There are useful variants of the Dominated Convergence Theorem in which the random sequence of interest is dominated by another integrable random sequence, as opposed to a single integrable random variable. Here is one such result. 

**Young’s Theorem.**

Let $x, x_1, x_2, \ldots$ be $\mathbb{R}$-valued random variables on a probability space $(X, \Sigma, p)$. Suppose there exist integrable $[0, \infty]$-valued random variables $y, y_1, y_2, \ldots$ on $(X, \Sigma, p)$ such that

$$|x_m| \leq \text{a.s. } y_m, \ m = 1, 2, \ldots, \quad y_m \to y \quad \text{and} \quad \mathbb{E}(y_m) \to \mathbb{E}(y).$$

Then,

$$x_m \to x \quad \text{implies} \quad \mathbb{E}(|x|) < \infty \quad \text{and} \quad \mathbb{E}(x_m) \to \mathbb{E}(x).$$

**Proof.** Assume that $x_m \to \text{a.s. } x$, and define $S := \{x_m \to x\} \cap \{y_m \to y\} \cap \{|x_1| \leq y_1\} \cap \{|x_2| \leq y_2\} \cap \cdots$. Clearly, $S \in \Sigma$ and $p(S) = 1$. Furthermore, $|x|1_S \leq y1_S$ so that $|x| \leq \text{a.s. } y$, whence $\mathbb{E}(|x|) \leq \mathbb{E}(y)$. As $y$ is integrable, therefore, $x$ is integrable as well. Next, observe that $(y_m + x_m)1_S \geq 0$ for each $m$, and we have $(y_m + x_m)1_S \to (y + x)1_S$. Moreover, $x_m$ and hence, $x_m1_S$, is integrable (for each $m$). Therefore, by Corollary 2.4 and Fatou’s Lemma,

$$\mathbb{E}(y) + \mathbb{E}(x) = \mathbb{E}(y1_S) + \mathbb{E}(x1_S) = \mathbb{E}((y + x)1_S) \leq \mathbb{E}(\lim\inf(y_m + x_m)1_S) = \lim\inf(\mathbb{E}(y_m) + \mathbb{E}(x_m)) = \mathbb{E}(y) + \lim\inf\mathbb{E}(x_m)$$

and hence $\mathbb{E}(x) \leq \lim\inf\mathbb{E}(x_m)$. As we would get $\mathbb{E}(x) \geq \lim\sup\mathbb{E}(x_m)$ by using the analogous reasoning with respect to the sequence $((y_m - x_m)1_S)$, it follows that $\mathbb{E}(x_m) \to \mathbb{E}(x)$, as we sought. □

---

17This theorem is often called Pratt’s Lemma (especially in the literature on mathematical statistics), but this is historically incorrect. William Young has proved it as early as in 1911, and yet this result seems to have been rediscovered many times after that. In particular, John Pratt has reproved it in 1960.
Integration of Almost Surely Defined Random Variables

Let \((X, \Sigma, p)\) be a probability space. Sometimes we need to deal with a map \(x\) that is defined on some proper \(\Sigma\)-measurable subset \(S\) of \(X\) with \(p(S) = 1\). If there exists a random variable \(y\) on \((X, \Sigma, p)\) such that \(y|_S = x\), we then say that such a map \(x\) is an almost surely defined random variable on \((X, \Sigma, p)\). Naturally, we define the expectation \(E(x)\) of \(x\) simply as that of \(y\), provided that \(E(y)\) exists. (By Corollary 2.4, this defines \(E(x)\) unambiguously.) When \(E(|x|) < \infty\), we refer to \(x\) as integrable.

**Example 2.4.** Let \(x\) and \(y\) be two \(\mathbb{R}\)-valued integrable random variables on a probability space \((X, \Sigma, p)\). Clearly, we cannot define \(x - y\) everywhere on \(X\), because both \(x\) and \(y\) may equal \(\infty\) at some \(\omega \in X\). Instead, we put \(S := \{x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}\), and define \(x - y\) as the real map on \(S\) that assigns to each \(\omega \in S\) the value \(x(\omega) - y(\omega)\). Notice that the integrability of \(x\) implies \(p\{x \in \mathbb{R}\} = 1\), and similarly, \(p\{y \in \mathbb{R}\} = 1\). Thus, \(p(S) = 1\). Now define \(z := x1_S - y1_S\). Given that \(z\) is the difference between two integrable random variables on \((X, \Sigma, p)\), we have \(z \in L^1(X, \Sigma, p)\). Moreover, \(z|_S = x - y\), so we conclude that \(x - y\) is an almost surely defined random variable on \((X, \Sigma, p)\). Similarly, \(|z| |_S = |x - y|\), so \(|x - y|\) is an almost surely defined random variable on \((X, \Sigma, p)\). Moreover, \(E(|x - y|) = E(|z|) \leq E(|x|) + E(|y|) < \infty\). Thus, \(x - y\) is an integrable almost surely defined random variable on \((X, \Sigma, p)\). \(\square\)

In the following set of exercises the expectations are taken relative to an arbitrarily fixed probability space \((X, \Sigma, p)\).

**Exercise 2.6.** Assume that \(X = [0, 1]\), and let \(x\) be an integrable random variable on \((X, \Sigma, p)\). Show that
\[
\int_X \omega^m x(\omega) p(d\omega) \to x(1)p\{1\}.
\]

**Exercise 2.7.** Assume that \(X = [0, 1]\), and let \(x\) be a nonnegative random variable on \((X, \Sigma, p)\). Show that
\[
E(x^m) \to \left\{\begin{array}{ll}
p\{x = 1\}, & \text{if } p\{x > 1\} = 0, \\
\infty, & \text{otherwise.}
\end{array}\right.
\]

**Exercise 2.8.** Let \((x_m)\) be a sequence of random variables on \((X, \Sigma, p)\). If \(\lim x_m \in \mathbb{R}^X\), \(\lim x_m \leq_{a.s.} 0\), and \(E(x_m) \to 0\), then \(E(|x_m|) \to 0\). True or false?

**Exercise 2.9.** Let \((x_m)\) be a sequence of integrable random variables on \((X, \Sigma, p)\) such that \(x_m \to x\) uniformly for some real map \(x\) on \(X\). Note that \(x\) is a random variable on \((X, \Sigma, p)\), and show that \(E(|x_m| - x) \to 0\). Conclude that \(E(x_m) \to E(x)\), provided that \(E(x)\) exists.

**Exercise 2.10.** Let \(x\) be an integrable random variable on \((X, \Sigma, p)\). Prove: If \((S_m)\) is a sequence in \(\Sigma\) with \(p(S_m) \to 0\), then \(E(x1_{S_m}) \to 0\).

**Exercise 2.11.** We now generalize the previous exercise. Let \(x\) be an integrable random variable on \((X, \Sigma, p)\), and define the real map \(\Phi\) on \(\Sigma\) by \(\Phi(S) := E(x1_S)\). Show that this map is continuous relative to the semimetric on \(\Sigma\) we introduced in Exercise B.2.10. Also show that if \(x\) is bounded, then \(\Phi\) is Lipschitz continuous relative to this semimetric.
Exercise 2.12. (Brezis-Lieb) Let $x, x_1, x_2, \ldots$ be nonnegative random variables on $(X, \Sigma, p)$ such that $x_m \to x$. Prove that

$$E(x_m) - E(x) - E(|x_m - x|) \to 0.$$ 

Exercise 2.13. Let $I$ be an open interval, and $x$ a random variable on $(X, \Sigma, p)$ such that

(i) $x(\cdot, t) \in L^1(X, \Sigma, p)$ for each $t \in I$;
(ii) $x(\omega, \cdot) \in C^1(I)$ for each $\omega \in X$; and
(iii) there is a $y \in L^1(X, \Sigma, p)$ with $y \geq a.s. \frac{\partial}{\partial t} x(\cdot, t)$ for each $t \in I$. Show that the real map $h$ on $I$, defined by (16), is differentiable, and

$$\frac{d}{dt} h(t) = \int_X \frac{\partial}{\partial t} x(\cdot, t) d\mu \quad \text{for each } t \in I.$$

Exercise 2.14. (Scheffé's Lemma) The conditions of Young's Theorem is not enough to guarantee that $E(|x_m - x|) \to 0$. But we can use this theorem to obtain simple sufficient conditions to ensure such convergence. Let $x, x_1, x_2, \ldots$ be nonnegative integrable random variables on $(X, \Sigma, p)$ such that $E(x_m) \to E(x)$ and $x_m \to a.s. x$. Scheffé's Lemma says that we must then have $E(|x_m - x|) \to 0$. Prove this!

Exercise 2.15. Let $x, x_1, x_2, \ldots$ be integrable random variables on $(X, \Sigma, p)$ such that $E(x_m) \to E(x)$. Prove: If $p \{x - x_m > \varepsilon\} \to 0$ for every $\varepsilon > 0$, then $E(|x_m - x|) \to 0$.

2.2 The Lebesgue Integral II

We have defined above the expectation of an $\mathbb{R}$-valued random variable by looking at the difference between the expectations of the positive and negative parts of that random variable. The same method allows us to extend the coverage of the Lebesgue integral to include such measurable functions as well. To say this precisely, let $(X, \Sigma, \mu)$ be a measure space, and $f : X \to \mathbb{R}$ a $\Sigma$-measurable map. The Lebesgue integral of $f$ (with respect to $\mu$) is defined as the (extended real) number

$$\int_X f d\mu := \int_X f^+ d\mu - \int_X f^- d\mu,$$

provided that the right-hand side of this expression is not of the $\infty - \infty$ form. (If the latter condition is not met, we say that the Lebesgue integral of $f$ (with respect to $\mu$) does not exist.) For any $S \in \Sigma$, we also define

$$\int_S f d\mu := \int_X f 1_S d\mu,$$

provided that the right-hand side of this expression exists. Finally, we say that $f$ is integrable (with respect to $\mu$) if $\int_X |f| d\mu < \infty$.

---

18 The Lebesgue integral was introduced as a generalization of the Riemann integral in 1901 by the great Henri Lebesgue (1875-1941). While it was evident that this generalization was quite novel, it is interesting that Lebesgue himself was apparently never confident about the promise of his work, as he was not too keen on generalizations. He is often quoted to say that “reduced to general theories, mathematics would be a beautiful form without content – it would quickly die.” At least in the case of his own generalized integration theory, however, we can safely say that his fears were entirely unwarranted.

35
The following set of exercises recite the properties of the expectation functional we considered above, this time for the Lebesgue integral. These are either deduced by imitating the corresponding arguments we used in the previous section, or by combining Proposition 1.4 with the results of that section. We will use them freely in the remainder of this text.

In the following set of exercises \((X, \Sigma, \mu)\) stands for an arbitrarily fixed measure space, while \(f\) and \(g\) are \(\Sigma\)-measurable functions from \(X\) into \(\mathbb{R}\).

Exercise 2.16. Prove: \(\int_X (af + g) d\mu = a \int_X f d\mu + \int_X g d\mu\) for any real number \(a\), provided that the Lebesgue integrals of \(f\) and \(g\) exist.

Exercise 2.17. Suppose that \(\int_X g d\mu\) exists and belongs to \((-\infty, \infty]\). Show that \(f \leq g\) implies \(\int_X f d\mu \geq \int_X g d\mu\).

Exercise 2.18. Prove: \(|\int_X f d\mu| \leq \int_X |f| d\mu\), provided that the Lebesgue integral of \(f\) exists.

Exercise 2.19. Suppose that the Lebesgue integral of \(f\) exists. Show that \(\int_S f d\mu\) exists for every \(S \in \Sigma\).

Exercise 2.20. Let \((X, 2^X, \mu)\) be a measure space where \(X\) is a countable set and \(\mu\) is the counting measure on \(2^X\). Prove that

\[
\int_X f d\mu = \sum_{\omega \in X} f(\omega)
\]

for any \(f \in \mathbb{R}^X\) which is integrable (with respect to \(\mu\)).

Exercise 2.21. Consider the measure space \((\mathbb{N}, 2^\mathbb{N}, \mu)\) where \(\mu\) is the counting measure on \(2^\mathbb{N}\). Let \(f\) stand for the real map on \(\mathbb{N}\) with \(f(i) := \frac{(-1)^{i-1}}{i}\). Show that the Lebesgue integral of \(f\) (with respect to \(\mu\)) does not exist.

Warning. The infinite series \(\sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i}\) converges, but the equation \(\int_{\mathbb{N}} \frac{(-1)^{i-1}}{i} \mu(di) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i}\) is false, because the left-hand side does not exist.

Exercise 2.22. In Exercise 2.9 we have seen that if a sequence of random variables (on a probability space) uniformly converges to a random variable, then the Lebesgue integrals of those functions converge to that of the limit random variable. Consider the measure space \((\mathbb{R}, B(\mathbb{R}), \ell)\) and the sequence \(\left(\frac{1}{n} 1_{[0,n]}\right)\) to show that this result is not valid in the case of infinite measure spaces.

Exercise 2.23. (The Lebesgue Convergence Theorem, Again) Let \(f_1, f_2, \ldots\) be a \(\Sigma\)-measurable \(\mathbb{R}\)-valued maps on \(X\) such that \(\mu\{\omega \in X : f_m(\omega) \to f(\omega)\\} = 0\). Show that

\[
\int_X f d\mu = \lim_{m \to \infty} \int_X f_m d\mu,
\]

if there is a \(\Sigma\)-measurable \(g : X \to [0, \infty]\) with \(\int_X g d\mu < \infty\) and \(g \geq \sup\{f, f_1, f_2, \ldots\}\).

2.3 The Change of Variables Formula

Changing Variables in Integration
Another important property of $E$ that we wish to investigate in this section concerns changing variables when computing the expectation of a transformation of a random variable. To see what we mean by this, take any random variable $x$ on some probability space $(X, \Sigma, p)$, and consider the Borel probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), p_x)$ induced by $x$. Now take any random variable $\varphi$ on the latter space. What is the expectation of $\varphi$? Applying the definition directly, the answer is $E_{p_x}(\varphi)$, that is, $\int_Y \varphi d p_x$. Alternatively, we could think of $\varphi \circ x$ as a random variable on the original space $(X, \Sigma, p)$, and compute $E_p(\varphi \circ x)$. While $E_{p_x}(\varphi)$ is the expectation of $\varphi$ with respect to the distribution of $x$, $E_p(\varphi \circ x)$ is computed by using the original probability measure $p$ but on the transformed random variable $\varphi \circ x$. One would surely hope that these two numbers are the same.

To illustrate, let us consider the experiment of throwing a single die, that is, the probability space $(X, 2^X, p)$ where $X := [6]$ and $p(S) := |S|/6$ for all $S \in 2^X$. Consider the random variable $x \in [2]^X$ which is defined as

$$x(\omega) := \begin{cases} 1, & \text{if } \omega \text{ is even} \\ 2, & \text{if } \omega \text{ is odd.} \end{cases}$$

In addition, define $\varphi : [2] \to \mathbb{R}$ by $\varphi(t) := t^2$, and observe that $\varphi$ is a $\{1, 4\}$-valued random variable on $([2], 2^2, p_x)$. Clearly, we have $E_{p_x}(\varphi) = (1)\frac{1}{2} + (4)\frac{1}{2} = \frac{5}{2}$; after all $\varphi$ takes value 1 with probability $\frac{1}{2}$ and 4 with probability $\frac{1}{2}$. On the other hand, $\varphi \circ x$ is a $\{1, 4\}$-valued random variable on $(X, 2^X, p)$ that takes value 1 if the outcome of the experiment is even and 4 if the outcome is odd. Thus, we have $E_p(\varphi \circ x) = \frac{5}{2}$.

The Change of Variables Formula

Intuition very much suggests that computing $E_p(\varphi \circ x)$ should in general be an alternative method of computing the expectation of $\varphi$. Fortunately, there are no surprises here (even in the more general context of the Lebesgue integral).

**The Change of Variables Formula.** Let $Y$ be a metric space, $x$ a $Y$-valued random variable on a probability space $(X, \Sigma, p)$, and $\varphi$ an $\mathbb{R}$-valued random variable on $(Y, \mathcal{B}(Y), p_x)$. If either $E_{p_x}(\varphi)$ or $E_p(\varphi \circ x)$ exists, then

$$\int_Y \varphi d p_x = \int_X (\varphi \circ x) d p.$$  \hspace{1cm} (17)

**Proof.** Assume first that $\varphi$ is simple. Then $\varphi(Y)$ is finite, and we have

$$\varphi = \sum_{a \in \varphi(Y)} a 1_{\{\varphi = a\}} \quad \text{and} \quad \varphi \circ x = \sum_{a \in \varphi(Y)} a 1_{\{\varphi \circ x = a\}}.$$  

(17) readily follows from these identities. Suppose next that $\varphi$ need not be simple but $\varphi(Y) \subseteq [0, \infty]$. In this case we use Proposition C.1.10 to find an increasing sequence
of nonnegative simple random variables on \((Y, \mathcal{B}(Y), \mathbf{p}_x)\) such that \(\varphi_m \uparrow \varphi\). But then \((\varphi_m \circ x)\) is a sequence of nonnegative simple random variables on \((X, \Sigma, \mathbf{p})\), and we have \(\varphi_m \circ x \uparrow \varphi \circ x\). So, by applying the Monotone Convergence Theorem 1 twice,
\[
\int_Y \varphi \, d\mathbf{p}_x = \lim \int_Y \varphi_m \, d\mathbf{p}_x = \lim \int_X (\varphi_m \circ x) \, d\mathbf{p} = \int_X (\varphi \circ x) \, d\mathbf{p}.
\]
This observation is readily extended to the case of an arbitrary \(\varphi\) such that either \(\mathbb{E}_{\mathbf{p}_x}(\varphi)\) or \(\mathbb{E}_{\mathbf{p}}(\varphi \circ x)\) exists, by applying what we now know to \(\varphi^+\) and \(\varphi^-\).

**Remark 2.1.** The Change of Variables Formula remains valid in the context of any measurable space. To wit, let \(Y\) be a metric space, and \(f\) a \(Y\)-valued random variable on a measurable space \((X, \Sigma, \mathbf{p})\). Let \(\mu_f : \mathcal{B}(Y) \to [0, \infty], \) defined by \(\mu_f(S) := \mu(f^{-1}(S))\), is a Borel probability measure on \(Y\) – this measure is sometimes called the **pushforward of \(\mu\) by \(f\)**. Moreover, for any \(\mathbb{R}\)-valued random variable \(\varphi\) on \((Y, \mathcal{B}(Y), \mu_f)\), we have \(\int_Y \varphi \, d\mu_f = \int_X (\varphi \circ f) \, d\mu\), provided that either the Lebesgue integral of \(\varphi\) with respect to \(\mu_f\) or that of \(\varphi \circ f\) with respect to \(\mu\) exists. The proof of this fact is identical to that we gave above, except that we should now use Exercise 1.17 instead of the Monotone Convergence Theorem 1.

**Examples**

**Example 2.5.** Let \(x\) be a nonnegative random variable on a probability space \((X, \Sigma, \mathbf{p})\) such that \(x(X)\) is countable. The Change of Variables Formula says that the expectation of \(x\) equals the Lebesgue integral of the identity function on \(x(X)\) with respect to the probability measure \(\mathbf{p}_x\). In view of Example 1.3, therefore,
\[
\mathbb{E}(x) = \sum_{a \in x(X)} a \mathbf{p}\{x = a\}.
\]
This formula is useful for computing the expectation of many interesting random variables.

**Example 2.6.** Take any \(p \in (0,1)\), and let \(x\) be an \(\mathbb{N}\)-valued random variable on a probability space \((X, \Sigma, \mathbf{p})\) such that \(\mathbf{p}\{x = i\} = p(1-p)^{i-1}\) for each \(i \in \mathbb{N}\). Such a random variable is said to have a **geometric distribution** with parameter \(p\). (Think of flipping a coin infinitely many times (as in Example B.3.2) with probability of heads coming up in any one toss being \(p\). We can then interpret \(x\) as the first time we observe heads in this process.) Using the formula obtained in Example 2.6, and recalling Example 1.1 of Appendix 1, we find that
\[
\mathbb{E}(x) = p + 2p(1-p) + 3p(1-p)^2 + \cdots = \frac{p}{1-p} \sum_{i=1}^{\infty} i(1-p)^i = \frac{p}{1-p} \frac{1}{(1 - (1-p))^2},
\]
that is, $\mathbb{E}(x) = \frac{1}{p}$. \hfill \Box

**Example 2.7.** Given a positive real number $\lambda > 0$, a $\mathbb{Z}_+\text{-valued}$ random variable $x$ on a probability space $(X, \Sigma, p)$ such that

$$p\{x = i\} = \frac{e^{-\lambda} \lambda^i}{i!}, \quad i = 0, 1, \ldots$$

is said to have a Poisson distribution with parameter $\lambda$. Now recall that $e^\lambda$ equals $1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \cdots$, so the formula obtained in Example 2.6 yields $\mathbb{E}(x) = \lambda$. (Check!) Similarly, $\mathbb{E}(x^2) = \lambda + \lambda^2$ and $\mathbb{V}(x) = \lambda$. (Exercise!) \hfill \Box

**Example 2.8.** For any real number $a$, recall that $[a]$ stands for the greatest integer less than or equal to $a$, that is, $[a] := \max\{k \in \mathbb{Z} : k \leq a\}$. Let $x$ be a nonnegative random variable on a probability space $(X, \Sigma, p)$. A surprisingly useful formula for computing the expectation of the $\mathbb{Z}_+\text{-valued}$ random variable $[x]$ is the following:

$$\mathbb{E}([x]) = \sum_{i=1}^{\infty} p\{x \geq i\}. \quad (18)$$

This is true because

$$\sum_{i=1}^{\infty} p\{x \geq i\} = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} p\{j + 1 > x \geq j\} = \sum_{j=1}^{\infty} j p\{[x] = j\} = \mathbb{E}([x])$$

where we owe the second equality to Corollary 1.10 of Appendix 1 and the third equality to the formula obtained in Example 2.6. Of course, if $x$ is itself $\mathbb{Z}_+\text{-valued}$, then $[x] = x$, so (18) tells us that

$$\mathbb{E}(x) = \sum_{i=1}^{\infty} p\{x \geq i\} \quad \text{for every } \mathbb{Z}_+\text{-valued random variable on } (X, \Sigma, p).$$

We will have plenty of occasion to make use of this observation. \hfill \Box

**Exercise 2.24.** Verify the claims made in Example 2.7.

**Exercise 2.25.** Let $x$ be a random variable that has a geometric distribution with parameter $p \in (0, 1)$ as in Example 2.6. Find $\mathbb{V}(x)$.

**Exercise 1.26.** The Riemann zeta function on $\{2, 3, \ldots\}$ is defined as $\zeta(k) := \sum_{i=1}^{\infty} i^{-k}$. It is not difficult to show that $\zeta$ is finite on its entire domain. (See Example 1.2 of Appendix 1.) It is more involved to compute the values of this function, but especially for even integers, power series analysis yields very nice results. In particular, it is known that $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$ and $\zeta(6) = \pi^6/945$. (You don’t have to prove any of these formulas.)

\textsuperscript{19}If you don’t know this (miraculous!) formula, that’s okay. I will prove it in Section E.1.6.
For any integer $k \geq 2$, an $\mathbb{N}$-valued random variable $x_k$ on a probability space $(X, \Sigma, \mathbb{P})$ such that

$$p\{x_k = i\} = \frac{i^{-k}}{\zeta(k)}, \quad i = 1, 2, \ldots$$

is said to have a zeta distribution with parameter $k$.

(a) Find the expected value of $x_k$ for each $k \geq 2$ in terms of the Riemann zeta function.
(b) Find the variance of $x_k$ for each $k \geq 3$ in terms of the Riemann zeta function.
(c) Compute $E(x_k^2)$ and $E(x_k^4)$.

Exercise 2.27. Prove: For any nonnegative random variable $x$ on a probability space $(X, \Sigma, \mathbb{P})$,

$$\sum_{i=1}^{\infty} p\{x > i\} \leq E(x) \leq \sum_{i=0}^{\infty} p\{x > i\}.$$

2.4 Uniform Integrability

Integrability of a random variable means simply that the expectation of this random variable is a real number. There is a useful way of extending this notion to the context of an arbitrary collection of random variables. While we will make use of this generalization only later in the text, we introduce it now, as all the ingredients needed for it are now in place.\(^\text{20}\)

Alternate Formulations of Integrability

Let us begin by making note of a few characterizations of integrability.

Lemma 2.7. For any random variable $x$ on a probability space $(X, \Sigma, \mathbb{P})$, the following statements are equivalent:

(a) $x$ is integrable,
(b) For every $\varepsilon > 0$, there is a real number $a > 0$ such that

$$\int_{\{|x| > a\}} |x| \, d\mathbb{P} < \varepsilon.$$

(c) For every $\varepsilon > 0$, there is a number $\delta \in (0, 1]$ such that

$$\int_S |x| \, d\mathbb{P} < \varepsilon \quad \text{for every } S \in \Sigma \text{ with } \mathbb{P}(S) < \delta,$$

Proof. Assume that $x$ is integrable, and fix an arbitrary $\varepsilon > 0$. For any positive integer $m$, put $x_m := -|x| 1_{\{|x| > m\}}$, and notice that $x_m \uparrow 0$ while $x_1$ is integrable (because so is $x$). By the Monotone Convergence Theorem 2, therefore, $E(x_m) \uparrow 0$, which means $E(|x| 1_{\{|x| > m\}}) \downarrow 0$. It follows that there is a real number $a > 0$ large enough that $E(|x| 1_{\{|x| > a\}}) < \varepsilon$. Conclusion: (a) implies (b).

\(^{20}\)To streamline the exposition, I will only develop the theory here in terms of probability spaces; but see Exercise 2.31.
Assume now that (b) holds. Take any \( \varepsilon > 0 \), and choose an \( a > 0 \) large enough that \( \mathbb{E}(|x| \mathbf{1}_{\{|x| > a\}}) < \frac{\varepsilon}{2} \). Then

\[
\int_S |x| \, dp = \int_{S \cap \{|x| > a\}} |x| \, dp + \int_{S \cap \{|x| \leq a\}} |x| \, dp < \frac{\varepsilon}{2} + a p(S)
\]

for any \( S \in \Sigma \). So, putting \( \delta := \varepsilon/2a \), we find that \( \mathbb{E}(|x| \mathbf{1}_S) < \varepsilon \) for every \( S \in \Sigma \) with \( p(S) < \delta \). Conclusion: (b) implies (c).

Finally, assume that (c) holds. Pick any \( \delta > 0 \) so that \( \mathbb{E}(|x| \mathbf{1}_S) < 1 \) for every \( S \in \Sigma \) with \( p(S) < \delta \). Since \( \{|x| > m\} \downarrow \{|x| = \infty\} = \emptyset \), we have \( p(|x| > m) \downarrow 0 \) by continuity of \( p \). Thus, there is an \( a > 0 \) with \( p(|x| > a) < \delta \). Taking \( S = \{|x| > a\} \) in (c), then,

\[
\int_X |x| \, dp = \int_{\{|x| > a\}} |x| \, dp + \int_{\{|x| \leq a\}} |x| \, dp < 1 + a p(|x| \leq a) < \infty.
\]

Conclusion: (c) implies (a).

**Example 2.9.** Here is a quick application of Lemma 2.7. Let \((X, \Sigma, p)\) be a probability space, and consider \( \Sigma \) as a semimetric space relative to the semimetric \( d_p \) introduced in Exercise B.2.10. Fix an integrable random variable \( x \) on \((X, \Sigma, p)\), and define \( \Phi : \Sigma \to \mathbb{R} \) by

\[
\Phi(S) := \int_S x \, dp.
\]

We have claimed in Exercise 2.11 that \( \Phi \) is continuous. While one may prove this directly, Lemma 2.7 yields readily a sharper result. Indeed, fix any \( \varepsilon > 0 \), and use part (c) of this lemma to choose a \( \delta > 0 \) such that \( \int_S |x| \, dp < \varepsilon \) for every \( S \in \Sigma \) with \( p(S) < \delta \). But then, for any \( A \) and \( B \) in \( \Sigma \) with \( p(A \Delta B) < \delta \), we have

\[
|\Phi(A) - \Phi(B)| \leq \int_{A \Delta B} |x| \, dp < \varepsilon.
\]

In view of the arbitrariness of \( S \) and \( \varepsilon \), we may thus conclude that \( \Phi \) is uniformly continuous.

**Exercise 2.28.** Let \( x \) be a nonnegative random variable on a probability space \((X, \Sigma, p)\). Prove that \( x \) is integrable iff \( \sup\{\mathbb{E}(\min\{x, m\}) : m = 1, 2, \ldots\} < \infty \).

**Application: Dominated Convergence Theorem, Again**

We have derived the Dominated Convergence Theorem above first by establishing the Monotone Convergence Theorem 2 and then using a suitable extension of Fatou’s Lemma for \( \mathbb{R} \)-valued random variables as well. For (c) to imply (a), however, we need at least that \( p\{|x| = \infty\} = 0 \), unless we know more about the structure of the underlying probability space.

---

\( ^{21} \) That (a) implies (b), and (b) implies (c), are true for \( \mathbb{R} \)-valued random variables as well. For (c) to imply (a), however, we need at least that \( p\{|x| = \infty\} = 0 \), unless we know more about the structure of the underlying probability space.
variables. We can also prove this theorem by a more direct argument without invoking Fatou’s Lemma. In fact, using Egorov’s Theorem (Exercise C.1.13) and Lemma 2.7 yields even a sharper result. Let us first put on record the following simple fact.

**Observation 1.1.** Let \( x, x_1, x_2, \ldots \) be integrable random variables on a probability space \((X, \Sigma, \mathbf{p})\), and take any \( T \in \Sigma \). If \( x_m \to x \) uniformly on \( T \), then \( \int_T |x_m - x| \, d\mathbf{p} \to 0 \).

**Proof.** For any \( \varepsilon > 0 \), the uniform convergence of \( (x_m) \) to \( x \) on \( T \) entails that there is a positive integer \( M \) such that \( |x_m(\omega) - x(\omega)| < \varepsilon \) for every \( \omega \in T \) and \( m \geq M \). But then \( \int_T |x_m - x| \, d\mathbf{p} < \varepsilon \) for each \( m \geq M \). \( \Box \)

Now, let \((X, \Sigma, \mathbf{p})\) and \( x, y, x_1, x_2, \ldots \) be exactly as in the Dominated Convergence Theorem, and assume \( x_m \to x \) a.s. Let us assume first that \( x, y, x_1, x_2, \ldots \) are all real-valued and \( x_m \to x \). We fix an arbitrary \( \varepsilon > 0 \), and use Lemma 2.7 to find a \( \delta > 0 \) such that \( \mathbb{E}(|y| 1_S) < \frac{\varepsilon}{2} \) for every \( S \in \Sigma \) with \( \mathbf{p}(S) < \delta \). We next use Egorov’s Theorem to find a \( T \in \Sigma \) such that \( \mathbf{p}(X \setminus T) < \delta \) and \( x_m \to x \) uniformly on \( T \). But then, as \( |x_m - x| \leq |x_m| + |x| \leq_{a.s.} 2|y| \), we have

\[
\mathbb{E}(|x_m - x|) \leq \int_T |x_m - x| \, d\mathbf{p} + \int_{X \setminus T} |x_m - x| \, d\mathbf{p} \\
\leq \int_T |x_m - x| \, d\mathbf{p} + 2 \int_{X \setminus T} |y| \, d\mathbf{p} \\
< \int_T |x_m - x| \, d\mathbf{p} + \varepsilon,
\]

so letting \( m \uparrow \infty \) yields \( \limsup \mathbb{E}(|x_m - x|) \leq \varepsilon \) in view of Observation 1.1. As \( \varepsilon > 0 \) is arbitrary here, we thus conclude that \( \mathbb{E}(|x_m - x|) \to 0 \). (This conclusion is stronger than that of the Dominated Convergence Theorem.) As each \( x_m \) and \( x \) are integrable here, we can use Propositions 2.2 and 2.1 to get \( |\mathbb{E}(x_m) - \mathbb{E}(x)| = |\mathbb{E}(x_m - x)| \leq \mathbb{E}(|x_m - x|) \), so it follows from what we have just shown that \( \mathbb{E}(x_m) \to \mathbb{E}(x) \).

Let us now drop the assumption that each \( x, y, x_1, x_2, \ldots \) is real-valued. In this general case, we put \( S := \{ y \in \mathbb{R} \} \cap \{ x_m \to x \} \), and note that \( S \in \Sigma \) and \( \mathbf{p}(S) = 1 \). But then applying what we have found above in the context of the random variables \( x1_S, y1_S, x_11_S, x_21_S, \ldots \), we may conclude that \( \mathbb{E}((x_m - x)1_S) \to 0 \) and \( \mathbb{E}(x_m1_S) \to \mathbb{E}(x1_S) \). Provided that we interpret \( x_m - x \) as defined on \( S \) — this makes \( x_m - x \) an almost surely defined random variable on \((X, \Sigma, \mathbf{p})\); recall Example 2.4 — we may conclude from this finding that \( \mathbb{E}(|x_m - x|) \to 0 \) and \( \mathbb{E}(x_m) \to \mathbb{E}(x) \), as we sought.

**Uniformly Integrable Collections of Random Variables**

Let \( \mathcal{X} \) be a nonempty collection of integrable random variables on a probability space \((X, \Sigma, \mathbf{p})\). By Lemma 2.7, for every \( \varepsilon > 0 \), there is function \( \varphi : \mathcal{X} \to \mathbb{R}_{++} \) such that

\[
\int_{\{|x| > \varphi(x)\}} |x| \, d\mathbf{p} < \varepsilon \quad \text{for every } x \in \mathcal{X}.
\]

If the map \( \varphi \) is bounded from above, that is, if \( \sup\{ \varphi(x) : x \in \mathcal{X} \} \) equals to a real number, say, \( a \), then we could replace \( \varphi \) in the expression above with the constant function on \( \mathcal{X} \) that takes value \( a \) everywhere. In this special case we say that the collection \( \mathcal{X} \) is uniformly integrable. (But this definition has to be modified in the context of infinite measure spaces; see Exercise 2.31) Put differently:

42
Definition. A collection $\mathcal{X}$ of random variables on a probability space $(X, \Sigma, p)$ is said to be uniformly integrable (with respect to $p$) if
\[
\lim_{a \to \infty} \sup \left\{ \int_{\{|x| > a\}} |x| \, dp : x \in \mathcal{X} \right\} = 0,
\]
that is, for every $\varepsilon > 0$, there is a real number $a > 0$ such that
\[
\int_{\{|x| > a\}} |x| \, dp < \varepsilon \quad \text{for each } x \in \mathcal{X}.
\]
(Note. This definition maintains that $\emptyset$ is uniformly integrable.) In turn, a sequence $(x_m)$ of random variables on $(X, \Sigma, p)$ is called uniformly integrable if $\{x_1, x_2, \ldots\}$ is uniformly integrable.\(^{22}\)

The importance of this concept will become clear later in the text. For now, we will only look at a few examples.

Example 2.10. An immediate consequence of Lemma 2.7 is that every finite set of integrable random variables on a probability space $(X, \Sigma, p)$ is uniformly integrable. More generally, any nonempty subset $\mathcal{X}$ of $L^1(X, \Sigma, p)$ is uniformly integrable, provided that $\mathcal{X} \setminus \mathcal{Y}$ is uniformly integrable for some finite subset $\mathcal{Y}$ of $\mathcal{X}$. \(\square\)

Example 2.11. Every uniformly bounded collection of random variables on a probability space $(X, \Sigma, p)$ is uniformly integrable. That is, if $\mathcal{X}$ is a nonempty set of random variables on $(X, \Sigma, p)$ with $\sup\{|x| : x \in \mathcal{X}\}$ being a (real-valued) random variable on $(X, \Sigma, p)$, then $\mathcal{X}$ is uniformly integrable. More generally, if
\[
\sup\{|x| : x \in \mathcal{X}\} \leq_{a.s.} y \quad \text{for some } y \in L^1(X, \Sigma, p),
\]
then $\mathcal{X}$ is uniformly integrable. This is again an easy consequence of Lemma 2.7. \(\square\)

Example 2.12. Let $(X, \Sigma, p)$ be a probability space, and $\mathcal{X}$ and $\mathcal{Y}$ two nonempty uniformly integrable sets of random variables on $(X, \Sigma, p)$. We wish to show that $\{x + y : (x, y) \in \mathcal{X} \times \mathcal{Y}\}$ is also uniformly integrable. The key observation in this regard is
\[
\{|x + y| > a\} \subseteq \{2 \max\{|x|, |y|\} > a\}
\]
for any $a > 0$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Thanks to this fact,
\[
\int_{\{|x+y|>a\}} |x + y| \, dp \leq \int_{\{\max\{|x|, |y|\} > a/2\}} 2 \max\{|x|, |y|\} \, dp \\
\leq 2 \int_{\{|x|>a/2\}} |x| \, dp + 2 \int_{\{|y|>a/2\}} |y| \, dp
\]
\(^{22}\)We can also characterize uniform integrability along the lines of part (c) of Lemma 2.7. I ask you to do this in Exercise 2.30 below.
for any $a > 0$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Our claim thus follows readily from this finding, and the uniform integrability of $\mathcal{X}$ and $\mathcal{Y}$. \hfill \square

In the following set of exercises (X, Σ, p) stands for an arbitrary probability space.

Exercise 2.29. Suppose $(x_m)$ and $(y_m)$ are two uniformly integrable sequences of random variables on $(X, \Sigma, p)$. Does $(x_m y_m)$ have to be uniformly integrable?

Exercise 2.30. Let $\mathcal{X}$ be a nonempty set of random variables on $(X, \Sigma, p)$. Prove that $\mathcal{X}$ is uniformly integrable $\iff$ (i) $\sup \{\mathbb{E}(|x|) : x \in \mathcal{X} \} < \infty$, and (ii) for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\sup \{\mathbb{E}(|x| 1_S) : x \in \mathcal{X} \} < \varepsilon$ for all $S \in \Sigma$ with $p(S) < \delta$.

Exercise 2.31. Let $(X, \Sigma, \mu)$ be a measure space and $(f_m)$ a sequence in $L^0(X, \Sigma)$. We say that $(f_m)$ is uniformly integrable (with respect to $\mu$) if

$$\sup_{x} \int_{X} |f_m| \, d\mu < \infty, \quad \lim_{a \to -\infty} \sup_{x} \int_{\{|f_m| > a\}} |f_m| \, d\mu = 0 \quad \text{and} \quad \lim_{\delta \to 0} \sup_{x} \int_{\{|f_m| \leq \delta\}} |f_m| \, d\mu = 0$$

where all the suprema are taken over the positive integers $m$.

(a) Show that this definition is the same as we gave in the text when $\mu$ is a probability measure.

(b) Prove: If $(f_m)$ is uniformly integrable, then for every $\varepsilon > 0$, there is a $\delta > 0$ such that $\sup_{S} \int_{S} |f_m| \, d\mu < \varepsilon$ for all $S \in \Sigma$ with $\mu(S) < \delta$.

(c) Give an example to show that the converse of the fact we found in part (b) is not true even if we are given that $\sup_{S} \int_{S} |f_m| \, d\mu$ is finite. (Compare with Exercise 2.30.)

Exercise 2.32. Let $x, x_1, x_2, \ldots$ be integrable random variables on $(X, \Sigma, p)$ with $\mathbb{E}(|x_m - x|) \to 0$. Prove that $(x_m)$ must be uniformly integrable.

Exercise 2.33. Let $(x_m)$ and $(y_m)$ be two sequences of random variables on $(X, \Sigma, p)$ such that $|x_m| \leq a.s. \ y_m$ for each $m$. Show that if $(y_m)$ is uniformly integrable, so is $(x_m)$.

Exercise 2.34. Let $\mathcal{X}$ be a nonempty set of random variables on $(X, \Sigma, p)$, and let $\mathcal{H}$ stand for the set of all Borel measurable functions $H : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t \to \infty} H(t)/t = \infty$.

(a) Show that if there is an $H \in \mathcal{H}$ such that $\sup \{\mathbb{E}(H(|x|)) : x \in \mathcal{X} \} < \infty$, then $\mathcal{X}$ is uniformly integrable. (Example. $\mathcal{X}$ is uniformly integrable if $\sup \{\mathbb{E}(|x|^{1+\alpha}) : x \in \mathcal{X} \} < \infty$ for some $\alpha > 0$.)

(b) A theorem of Charles de la Vallée-Poussin says that the converse of the previous statement is true as well. To prove this, suppose $\mathcal{X}$ is uniformly integrable, and use this to find a sequence $(a_m)$ of positive real numbers such that $a_m \uparrow \infty$ and $\sup \{\mathbb{E}((|x| - a_m)^+) : x \in \mathcal{X} \} < 2^{-m}$ for each $m$. Now define the self-map $H$ on $\mathbb{R}_+$ by $H(t) := \sum_{k=0}^{\infty} \max\{|t| - a_k, 0\}$. Verify that $H$ is a convex function that belongs to $\mathcal{H}$, and then use the Monotone Convergence Theorem 1 to show that $\mathbb{E}(H(|x|)) \leq 1$ for each $x \in \mathcal{X}$.

Application. Let $(x_m)$ be a sequence of random variables on $(X, \Sigma, p)$ such that $\mathbb{E}(x_m) = 0$ and $\forall(x_m) = 1$ for each $m$. Then, $(x_m)$ is uniformly integrable (because $\mathbb{E}(x_m^2) = \sup \forall(x_m) = 1$).

Exercise 2.34. Let $\tau$ be a self-map on $X$ such that $\tau^{-1}(S) \in \Sigma$ and $p(\tau^{-1}(S)) = p(S)$ for every $S \in \Sigma$. For any $x \in L^1(X, \Sigma, p)$, prove that $\{x, x \circ \tau, x \circ \tau^2, \ldots\}$ and $\{\frac{1}{m} \sum_{i \in [m]} x \circ \tau^{i-1} : m \in \mathbb{N}\}$ are uniformly integrable.
2.5 Expectation of Banach Space-Valued Random Variables

There is a bit of a disconnect between the previous chapter and the present one. In Chapter B, we were able to carry out our initial investigation of probability theory without restricting our attention to the real line. In particular, we have allowed a random variable to take values in an arbitrarily given metric space \( X \) (but as Proposition C.1.4 hints at, a useful theory arises from such objects only when \( X \) is separable). By contrast, we have developed the Lebesgue integral in this chapter only for \( \mathbb{R} \)-valued random variables. Our goal in the present (optional) section is to see how one may relax the hypothesis of \( \mathbb{R} \)-valuedness in this development.

There is an immediate concern. The notion of expectation is an integration operation, and “linearity” is at the heart of this operation. As linearity is not meaningful in the context of an arbitrary metric space, therefore, it is not possible to extend our current development to the context of \( Y \)-valued random variables, unless we impose some linear structure on \( Y \). We can do this by taking \( Y \) as what is called a metric linear space, but such spaces almost always appear in probability theory in the form of normed linear spaces. Consequently, \( Y \) will be a normed linear space for us throughout this section. Moreover, to be able to develop a sound theory, we will often take \( Y \) as a Banach space (which will be separable in most cases). Our task is, then, to extend the theory of the Lebesgue integral to the context of Banach space-valued random variables.\(^{23}\)

Before we proceed, let us agree to denote the collection of all \( Y \)-valued random variables on a measurable space \((X, \Sigma)\) by \( \mathcal{L}^{0,Y}(X, \Sigma) \). (Thus, \( \mathcal{L}^{0,\mathbb{R}}(X, \Sigma) \) is none other than \( \mathcal{L}^{0}(X, \Sigma) \).) It is an easy exercise to show that \( x \in \mathcal{L}^{0,Y}(X, \Sigma) \) implies \( ax \in \mathcal{L}^{0,Y}(X, \Sigma) \) for any real number \( a \). Moreover, we know from Proposition C.1.4 that \( x + y \in \mathcal{L}^{0,Y}(X, \Sigma) \) for any \( x, y \in \mathcal{L}^{0}(X, \Sigma) \). Thus, \( \mathcal{L}^{0,Y}(X, \Sigma) \) is a linear space under the usual (pointwise) addition and scalar multiplication operations.

**The Pettis Measurability Theorem**

Let \( Y \) be a normed linear space (whose norm we denote by \( \| \cdot \|_Y \)), and \( x \) a \( Y \)-valued random variable on \((X, \Sigma)\). Unlike how things are when \( Y = \mathbb{R} \), the Borel measurability of \( x \) does not ensure that we can approximate \( x \) (pointwise) by means of simple \( Y \)-valued random variables on \((X, \Sigma)\), and this indeed makes a poor start for what we have in mind. But a famous theorem of measure theory, which was proved by Billy Pettis in 1938, says that all goes well when \( Y \) is a separable Banach space.

**The Pettis Measurability Theorem.** Let \( Y \) be a separable Banach space and \((X, \Sigma)\) a measurable space. For any map \( x : X \to Y \), the following are equivalent:

(a) \( x \) is a \( Y \)-valued random variable on \((X, \Sigma)\),

(b) \( L \circ x \) is a random variable on \((X, \Sigma)\) for every \( L \in Y^* \),\(^{24}\)

(c) There is a sequence \((x_m)\) of simple \( Y \)-valued random variables on \((X, \Sigma)\) such that \( x_m \to x \).

**Proof.** As it is continuous, any \( L \) in \( Y^* \) is a random variable on \((Y, \mathcal{B}(Y))\). As the composition of two random variables is a random variable, therefore, the fact that (a) implies (b) is straightforward.

\(^{23}\)I will only provide a precursory introduction to this topic here, but I will develop it further (in the context of probability theory) at various points in the text. If you wish to have a look at a more systematic development of this topic (which is bound to be more advanced), have a look at Chapter 2 of Diestel and Uhl (1977) and/or Chapter 6 of Lang (1996).

\(^{24}\)Reminder. \( Y^* \) is the set of all continuous linear functionals on \( Y \). When endowed with the usual operations of scalar multiplication and addition, and the operator norm, this set becomes itself a Banach space, which is called the dual of \( Y \). I denote the operator norm on \( Y^* \) by \( \| \cdot \|^* \), that is, \( \|L\|^* := \sup \{ |L(\nu)| : \nu \in Y \text{ and } \|\nu\|_Y = 1 \} \). (See Section 2.2 of Appendix 3.)

I will assume throughout this section that you are familiar with the basic theory of normed linear spaces, at least insofar as this theory is covered in Section 2 of Appendix 3.
Besides, we already know from Proposition C.1.5 that (c) implies (a). It remains to prove that (b) implies (c). Let us then assume that (b) holds. As $Y$ is separable, we can find a countably infinite subset of $Y^*$, which we enumerate as $\{L_1, L_2, \ldots\}$, such that $\|L_m\|^* = 1$ for each $m \in \mathbb{N}$, and $\|\nu\|_Y = \sup\{|L_m(\nu)| : m \in \mathbb{N}\}$ for each $\nu \in Y$. (See Corollary 2.9 of Appendix 3.) By hypothesis, $L_m \circ x$ is a random variable on $(X, \Sigma)$ for each $m$. But then, for any $\nu \in Y$ and positive integer $m$, $|L_m \circ x - L_m(\nu)|$ is a random variable on $(X, \Sigma)$, so the real map $\|x - \nu\|_Y$, being equal to $\sup\{|L_m \circ x - L_m(\nu)| : m \in \mathbb{N}\}$, is a random variable on $(X, \Sigma)$. Conclusion: For every $\nu \in Y$, the map $F_{\nu} : X \to \mathbb{R}_+$, defined by,

$$F_{\nu}(x) := \|x(\omega) - \nu\|_Y,$$

is a random variable on $(X, \Sigma)$.

Let us now pick a sequence $(\nu_m)$ in $Y$ such that $\{\nu_1, \nu_2, \ldots\}$ is dense in $Y$. For each $\nu \in Y$ and $m \in \mathbb{N}$, let $k(m, \nu)$ be the smallest integer $k$ in $[m]$ that minimizes $\|\nu - \nu_k\|_Y$. Since $\{\nu_1, \nu_2, \ldots\}$ is dense in $Y$, we have

$$\|\nu - \nu_{k(m, \nu)}\|_Y \to 0 \quad \text{for every } \nu \in Y$$

as $m \uparrow \infty$. For each $m$, we define the map $x_m : X \to Y$ by $x_m(\omega) := \nu_{k(m, x(\omega))}$. Then,

$$\|x(\omega) - x_m(\omega)\|_Y \to 0 \quad \text{for every } \omega \in X,$$

that is, $x_m \to x$, as $m \uparrow \infty$. We may thus conclude that (c) holds, if we can show that each $x_m$ is simple random variable on $(X, \Sigma)$. To this end, fix any positive integer $m$. Clearly, $x_m(X)$ is contained in $\{\nu_1, \ldots, \nu_m\}$. Moreover, for every $k \in [m]$, the definition of $k(m, \nu)$ entails that $x_m(\omega) = \nu_k$ if $\|x(\omega) - \nu_k\|_Y$ is the minimum number in $\{F_{\nu_1}(\omega), \ldots, F_{\nu_m}(\omega)\}$ and for no $l \in [m]$ with $l < k$, the number $\|x(\omega) - \nu_l\|_Y$ is the minimum of $\{F_{\nu_1}(\omega), \ldots, F_{\nu_m}(\omega)\}$. In other words, $\{x_m = \nu_k\} = \{F_{\nu_k} = \min\{F_{\nu_1}, \ldots, F_{\nu_m}\}\} \cap \bigcap_{l \in [k-1]} \{F_{\nu_l} > \min\{F_{\nu_1}, \ldots, F_{\nu_m}\}\}$

for any $k \in \{2, \ldots, m\}$. As $F_{\nu_1}, \ldots, F_{\nu_m} \in \mathcal{L}^0(X, \Sigma)$, therefore, $\{x = \nu_k\} \in \Sigma$ for each $k \in [m]$, and it follows that $x_m$ is a simple $Y$-valued random variable on $(X, \Sigma)$. Conclusion: (b) implies (c).

**Remark 2.2.** Let $Y$ be a Banach space. In functional analysis, a map $x : X \to Y$ that satisfies (b) of the theorem above is said to be *weakly $\Sigma$-measurable*. In turn, those maps in $X^Y$ that are pointwise limits of simple $Y$-valued random variables on $(X, \Sigma)$ are called *strongly $\Sigma$-measurable*. Proposition C.1.5 shows that strong measurability always implies weak measurability, but it can be shown that the converse is false in general. And yet the Pettis Measurability Theorem tells us that these are equivalent notions when $Y$ is separable. More generally, these notions coincide for any $x : X \to Y$ such that $x(X)$ is contained within a separable closed subspace of $Y$. □

**Bochner Integrability**

Let $Y$ be a Banach space and $(X, \Sigma, p)$ a probability space. For any simple $Y$-valued random variable $x$ on $(X, \Sigma, p)$, we define the expectation of $x$ as

$$E(x) := \sum_{\nu \in x(X)} \nu p\{x = \nu\}.$$

(Notice that $E(x)$ is a vector in $Y$.) Obviously, this definition reduces to the one we gave in Section 1.1 for $Y = \mathbb{R}$. Also, just as in that case (Exercise 1.1), we have

$$E \left( \sum_{i=1}^n \nu_i 1_{S_i} \right) = \sum_{i=1}^n \nu_i p(S_i)$$

46
for any positive integer \( n \), partition \( \{ S_1, \ldots, S_n \} \) of \( X \), and \( \nu_1, \ldots, \nu_n \in Y \). Moreover, it is readily checked that
\[
\| \mathbb{E}(x) \|_Y \leq \mathbb{E}(\| x \|_Y) \tag{19}
\]
and
\[
\mathbb{E}(x \mp y) = \mathbb{E}(x) \mp \mathbb{E}(y) \tag{20}
\]
for any simple \( Y \)-valued random variables \( x \) and \( y \) on \( (X, \Sigma, p) \).

We would like to extend this definition to the context of non-simple \( Y \)-valued random variables on \( (X, \Sigma, p) \). We cannot quite mimic what we did in Section 1.2 in this regard, because there is no natural order structure on \( Y \). Nevertheless, our development of the Lebesgue integral suggests that if we can approximate a given \( Y \)-valued random variable \( x \) on \( (X, \Sigma, p) \) with simple \( Y \)-valued random variables to any desired degree of accuracy, it may be a good idea to define \( \mathbb{E}(x) \) as the limit of the expectations of these simple random variables. But, of course, what we mean by “approximation” is detrimental here. Loosely speaking, we will take this to mean in the sense of both pointwise convergence and convergence in the mean. As was shown by Salomon Bochner in 1932, this particular notion of approximation leads to an integration theory which extends Lebesgue’s theory quite well.

**Definition.** Let \( Y \) be a Banach space and \( x \) a \( Y \)-valued random variable on a probability space \( (X, \Sigma, p) \). We say that \( x \) is **Bochner integrable** (with respect to \( p \)) if there is a sequence \((x_m)\) of simple \( Y \)-valued random variables on \( (X, \Sigma, p) \) such that

\[
x_m \to x \quad \text{and} \quad \int_X \|x_m - x\|_Y \, dp \to 0. \tag{21}
\]

Before we see how to utilize this notion to define an integral for \( Y \)-valued random variables, there are two issues to deal with. First, it is not obvious if Bochner integrability reduces to our usual notion of integrability when \( Y = \mathbb{R} \). Second, the definition of Bochner integrability is not quite friendly when it comes to checking whether or not a given Banach space-valued random variable possesses this property. The following proposition puts both of these issues to rest at one stroke.

**Proposition 2.8.** (Bochner) Let \( Y \) be a Banach space and \( x \) a \( Y \)-valued random variable on \( (X, \Sigma, p) \). If \( x \) is Bochner integrable, then

\[
\int_X \| x \|_Y \, dp < \infty.
\]

If \( Y \) is separable, the converse of this implication is also true, that is, in that case, \( x \) is Bochner integrable if, and only if, \( \| x \|_Y \) is integrable.

**Proof.** Suppose \( x \) is Bochner integrable, and \((x_m)\) is a sequence of simple \( Y \)-valued random variables on \( (X, \Sigma, p) \) such that (21) holds. Then, there is a positive integer \( M \) large enough that \( \int_X \| x_M - x \|_Y \, dp \leq 1 \). Consequently, by the subadditivity of \( \| \cdot \|_Y \),

\[
\int_X \| x \|_Y \, dp \leq \int_X \| x_M - x \|_Y \, dp + \int_X \| x_M \|_Y \, dp \leq 1 + \int_X \| x_M \|_Y \, dp < \infty.
\]

Conversely, suppose \( Y \) is separable and \( \| x \|_Y \) is integrable. By the Pettis Measurability Theorem, there is a sequence \((y_m)\) of simple \( Y \)-valued random variables on \( (X, \Sigma, p) \) such that \( y_m \to x \). Now put \( S_m := \{ x \neq 0 \text{ and } \| y_m \|_Y \leq 2 \| x \|_Y \} \), and define \( x_m := y_m 1_{S_m} \) for each \( m \). (Here \( 0 \) stands for the origin of \( Y \).) For any \( \omega \in X \), as \( y_m(\omega) \to x(\omega) \) and \( \| \cdot \|_Y \) is continuous, either \( \omega \) belongs to

\[\text{47}\]
all but finitely many of \( S_m \)'s, or \( x(\omega) = 0 \). It follows that \( x_m(\omega) \to x(\omega) \) for any \( \omega \in X \), that is, \( x_m \to x \). Moreover, \( \|x_m\|_Y \leq 2\|x\|_Y \) for each \( m \), so \( \|x_m - x\|_Y \leq \|x_m\|_Y + \|x\|_Y \leq 3\|x\|_Y \), and hence, given that \( \|x\|_Y \) is integrable, we may apply the Dominated Convergence Theorem to find
\[
\int_X \|x_m - x\|_Y \, dp \to 0.
\]

In what follows, for any Banach space \( Y \), we denote the collection of all Bochner integrable \( Y \)-valued random variables on a probability space \((X, \Sigma, p)\) by \( B\mathcal{L}^{1,Y}(X, \Sigma, p) \). In view of Proposition 2.8, therefore,
\[
B\mathcal{L}^{1,Y}(X, \Sigma, p) = \left\{ x \in \mathcal{L}^{0,Y}(X, \Sigma) : \int_X \|x\|_Y \, dp < \infty \right\}.
\]

Using the subadditivity of \( \| \cdot \|_Y \), one can easily show that \( B\mathcal{L}^{1,Y}(X, \Sigma, p) \) is a linear subspace of \( \mathcal{L}^{0,Y}(X, \Sigma) \) under the usual operations of addition and scalar multiplication.

**The Bochner Integral**

We now turn to the main task at hand, and use the notion of Bochner integrability to define the expectation of Banach space-valued random variables by means of approximation. Put precisely, we define the expectation of a Bochner integrable \( Y \)-valued random variable on \((X, \Sigma, p)\) as the vector \( \mathbb{E}(x) \) in the Banach space \( Y \) with \( \mathbb{E}(x) := \lim \mathbb{E}(x_m) \), where \((x_m)\) is any sequence of simple \( Y \)-valued random variables on \((X, \Sigma, p)\) such that (21) holds. In functional analysis, \( \mathbb{E}(x) \) is called the **Bochner integral** of \( x \) with respect to \( p \), and is denoted by \( B\int_X x \, dp \). As you would expect, we also set
\[
B\int_S x \, dp := \mathbb{E}(x1_S) \quad \text{for any } S \in \Sigma,
\]
where \( 1_S \) is the indicator function of \( S \) on \( X \).

We should of course check that \( \mathbb{E}(x) \) is well-defined here. That is, we need to show that (21) ensures that \( \mathbb{E}(x_m) \) converges in \( Y \), and also that this limit does not depend on which sequence of simple \( Y \)-valued random variables is used in the approximation (so long as that sequence satisfies (21)). The “Banachness” of \( Y \) ensures that this is indeed the case, as the following claim demonstrates.

**Claim.** Let \((x_m)\) and \((y_m)\) be two sequences of simple \( Y \)-valued random variables on \((X, \Sigma, p)\) such that \( \mathbb{E}(\|x_m - x\|_Y) \to 0 \) and \( \mathbb{E}(\|y_m - x\|_Y) \to 0 \) for some \( Y \)-valued random variable \( x \) on \((X, \Sigma, p)\). Then, \( (\mathbb{E}(x_m)) \) and \( (\mathbb{E}(y_m)) \) converge to the same vector in \( Y \).

**Proof.** For any positive integers \( k \) and \( l \), using (20), (19) and the subadditivity of \( \| \cdot \|_Y \) yields
\[
\|\mathbb{E}(x_k) - \mathbb{E}(x_l)\|_Y \leq \mathbb{E}(\|x_k - x_l\|_Y) \leq \mathbb{E}(\|x_k - x\|_Y) + \mathbb{E}(\|x - x_l\|_Y),
\]
and it follows that \( (\mathbb{E}(x_m)) \) is Cauchy in \( Y \), and similarly for \( (\mathbb{E}(y_m)) \). As \( Y \) is complete, therefore, both \( (\mathbb{E}(x_m)) \) and \( (\mathbb{E}(y_m)) \) converge in \( Y \). Moreover,
\[
\|\mathbb{E}(x_m) - \mathbb{E}(y_m)\|_Y \leq \mathbb{E}(\|x_m - x\|_Y) + \mathbb{E}(\|x - y_m\|_Y),
\]
so the limits of these sequences in \( Y \) must be the same. \( \|
\]

Unsurprisingly, the Bochner integral and the Lebesgue integral coincide in the context of integrable (real-valued) random variables (where, naturally, we use the absolute value function to norm \( \mathbb{R} \)). Indeed, if \( x \in \mathcal{L}^1(X, \Sigma, p) \), then Proposition 2.8 says that \( x \) is Bochner integrable (as an \( \mathbb{R} \)-valued random variable), so there is a sequence \((x_m)\) of simple random variables on \((X, \Sigma, p)\) such that (21) holds, and we have \( B\int_X x_m \, dp \to \int_X x \, dp \). Besides, (21) implies \( \int_X x_m \, dp \to \int_X x \, dp \) while, by definition, the Bochner and Lebesgue integrals coincide for simple random variables. Thus:
Let $Y$ be a Banach space and $x$ a Bochner integrable $Y$-valued random variable on $(X, \Sigma, \mathfrak{p})$. Then,

$$L(\mathbb{E}(x)) = \int_X (L \circ x) d\mathfrak{p} \quad \text{for every } L \in Y^*,$$

and $\mathbb{E}(x)$ is the only vector in $Y$ with this property.

**Proof.** An easy consequence of the Hahn-Banach Theorem is that for any two distinct vectors in $Y$ there is a continuous linear functional on $Y$ which assumes distinct values at those vectors. (See Corollary 2.8 of Appendix 3.) The uniqueness part of our assertion follows immediately from this observation. On the other hand, its existence part is easily seen to be valid if $x$ is simple on account of the linearity of the maps in $Y^*$. To settle the general case, pick any $L \in Y^*$ and any sequence $(x_m)$ of simple $Y$-valued random variables on $(X, \Sigma, \mathfrak{p})$ with (21), and note that

$$L(\mathbb{E}(x)) = L(\lim \mathbb{E}(x_m)) = \lim L(\mathbb{E}(x_m)) = \lim \int_X (L \circ x_m) d\mathfrak{p} \quad (22)$$

since $L$ is continuous and our claim is true for simple $Y$-valued random variables. Now put $y_m := \|L\|^* \|x_m\|_Y$ for each $m$, and $y := \|L\|^* \|x\|_Y$. By definition of the operator norm, we have $|L \circ x_m| \leq y_m$ and $|L \circ x| \leq y$ while $y_m \to y$ by continuity of $\| \cdot \|_Y$. Besides, each $y_m$ is a simple random variable, and hence, integrable, while $y$ is integrable by Proposition 2.8. Moreover,

$$|\mathbb{E}(y_m) - \mathbb{E}(y)| \leq \|L\|^* \int_X \|x_m\|_Y - \|x\|_Y \, d\mathfrak{p} \leq \|L\|^* \int_X \|x_m - x\|_Y \, d\mathfrak{p} \to 0$$

by the choice of $(x_m)$. We may then apply Young’s Theorem to conclude that

$$\lim \int_X (L \circ x_m) d\mathfrak{p} = \int_X \lim (L \circ x_m) d\mathfrak{p} = \int_X (L \circ x) d\mathfrak{p}$$

where the second equality follows from the continuity of $L$. Combining (22) with this finding completes our proof. 

Recall that the expectation of a random variable is some sort of a weighted average of the values of that random variable. In particular, this expectation has to fall between the infimum and supremum values of the random variable. A similar observation also holds for the expectation operator defined through the Bochner integral. In particular, this operator always takes values within the closed convex hull of the values of the random variable at hand. This is an easy consequence of Theorem 2.9.

**Proposition 2.10.** Let $Y$ be a Banach space and $x$ a Bochner integrable $Y$-valued random variable on $(X, \Sigma, \mathfrak{p})$. Then, $\mathbb{E}(x) \in \overline{co}(x(X))$.

**Proof.** Suppose our assertion is false. Then, by the Separating Hyperplane Theorem (Corollary 2.11 of Appendix 3), there is an $L \in Y^*$ such that $\mathbb{E}(\mathbb{E}(x)) > \sup \mathbb{E}(\overline{co}(x(X)))$. But the right-hand side of this inequality is larger than $L \circ x$ everywhere, so $\sup \mathbb{E}(\overline{co}(x(X))) \geq \mathbb{E}(L \circ x)$. Combining these inequalities yields a contradiction to Theorem 2.9.

}$
In the following set of exercises \((X, \Sigma, p)\) stands for an arbitrary probability space and \(Y\) a Banach space.

**Exercise 2.35.** Show that the Bochner integral is a linear operator on \(B\mathcal{L}^{1,Y}(X, \Sigma, p)\) such that
\[
\left\| \int_X x \, dp \right\|_Y \leq \int_X \|x\|_Y \, dp \quad \text{for any } x \in B\mathcal{L}^{1,Y}(X, \Sigma, p).
\]

**Exercise 2.36.** (Dominated Convergence Theorem for the Bochner Integral) Take any \(Y\)-valued random variables \(x_1, x_2, \ldots \) on \((X, \Sigma, p)\) and suppose \(x_m \to_{a.s.} x\) for some \(x \in Y^X\). Prove: If there is a \(y \in L^1(X, \Sigma, p)\) such that \(\|x_m\| \leq a.s. \ y\) for each \(m\), then \(\lim \mathbb{E}(\|x_m - x\|_Y) = 0\), and hence \(\lim \mathbb{E}(x_m) = \mathbb{E}(x)\).

**Exercise 2.37.** (Change of Variables Formula for the Bochner Integral) Assume that \(Y\) is separable, and take any metric space \(Z\). Let \(x\) be a \(Z\)-valued random variable on \((X, \Sigma, p)\), and \(\varphi\) a \(Y\)-valued random variable on \((Z, B(Z), p_x)\). Show that \(\varphi\) is Bochner integrable with respect to \(p_x\) iff \(\varphi \circ x\) is Bochner integrable with respect to \(p\), and when either (hence both) of these conditions hold,
\[
B \int_Z \varphi \, dp_x = B \int_X (\varphi \circ x) \, dp.
\]

**Exercise 2.38.** Take any compact metric space \(Z\), and note that \(C(Z)\) is a separable Banach space. Let \(x\) be a Bochner integrable \(C(Z)\)-valued random variable on \((X, \Sigma, p)\). For any \(\nu \in Z\), define \(x_\nu : X \to \mathbb{R}\) by \(x_\nu(\omega) := x(\omega)(\nu)\), and prove that
\[
\left( B \int_X x \, dp \right)(\nu) = \int_X x_\nu \, dp.
\]

**Exercise 2.39.** Let \(Z\) be a Banach space and \(L\) a continuous linear operator from \(Y\) into \(Z\). For any Bochner integrable \(Y\)-valued random variable \(x\) on \((X, \Sigma, p)\), show that \(L(\mathbb{E}(x)) = \mathbb{E}(L(x))\).

**Exercise 2.40.** A \(Y\)-valued random variable is said to be **Pettis integrable** if \(\mathbb{E}(\|L \circ x\|) < \infty\) for each \(L \in Y^*\). Show that Bochner integrability implies Pettis integrability. Also show that Bochner and Pettis integrability is the same for simple Banach space-valued random variables.

**Exercise 2.41.** Recall that \(c_0\) stands for the normed linear space of all real sequences that converge to zero (relative to the sup-norm); this is a separable Banach space. Consider the map \(x : [0, 1] \to c_0\) where \(x(\omega)\) is the real sequence whose \(n\)th term is \(m\) if \(\omega \in \left(\frac{1}{m+1}, \frac{1}{m}\right]\), and 0 otherwise. (For instance, \(x_1(\frac{2}{3}) = (0, 2, 0, \ldots)\).) This map is a \(c_0\)-valued random variable on \(([0, 1], \mathcal{B}[0, 1], \ell)\).

(a) Show that \(\|x\|_{c_0} = \sum_{m=1}^{\infty} \frac{1}{m+1} \), and use this to conclude that \(x\) is not Bocher integrable.

(b) We have proved in Example 2.7 of Appendix 3 that for any \(L \in c_0^*\), there is a sequence \((c_m) \in \ell^1\) such that \(L(a_1, a_2, \ldots) = \sum_{m=1}^{\infty} c_m a_m\) for each \((a_m) \in c_0\). Use this fact to prove that \(L(\frac{1}{2}, \frac{1}{3}, \ldots) = \mathbb{E}(L(x))\) for each \(L \in c_0^*\), and conclude that \(x\) is Pettis integrable.

### 3 Application: First-Order Stochastic Dominance

As an application of what we have accomplished thus far, we now make our first pass at the theory of stochastic dominance. This theory aims at formulating comparison methods for stochastic models, and, as we will see later, it is widely used in the
contexts of economics of uncertainty and welfare.\textsuperscript{26} In this section, we will introduce the part of the theory that provides a rigorous method of thinking about a random variable being \textit{stochastically larger} than another one in a way that goes significantly beyond the overly simple procedure of comparing the expectations of these variables.

3.1 The Univariate Case

When would we say that a random variable $x$ is larger than another random variable $y$? Put a bit more precisely, when would you think that, in all likelihood, the realization of $x$ is bound to exceed that of $y$? Since $x$ and $y$ need not be defined on the same probability space, we may not simply require $x \geq y$ (or $x \geq_{a.s.} y$) to this end. Of course, we can ask for $\mathbb{E}(x) \geq \mathbb{E}(y)$, but that alone is a long way from asserting that $x$ is larger than $y$ in unambiguous terms. It seems what we need is some form of dominance of $x$ over $y$. Enter, the first-order stochastic dominance.

\textbf{Definition.} Let $x$ and $y$ be two random variables. We say that $x$ \textbf{first-order stochastically dominates} $y$, and write $x \succ_{\text{FSD}} y$, if

$$
\mathbb{E}(u \circ x) \geq \mathbb{E}(u \circ y) \text{ for every increasing } u \in \mathcal{B}(\mathbb{R}).
$$

When at least one of these inequalities is strict, we write $x \succ_{\text{FSD}} y$.

Clearly, the first-order stochastic dominance is a preorder when applied on a prespecified set of random variables. However, it is not a partial order in general. Indeed, if $x$ and $y$ are two distinct but almost surely equal random variables on a common probability space, then we are bound to have $x \succ_{\text{FSD}} y \succ_{\text{FSD}} x$, that is, $\succ_{\text{FSD}}$ is not antisymmetric in general.

Before examining the structure of first-order stochastic dominance further, we go through a few examples that illustrate how this preorder arises in economic analysis.

\textbf{Example 3.1.} Consider an individual who is contemplating about choosing between the gambles (random variables) $x$, $y$ and $z$, where $x$ may pay out 1, 2 or 3 dollars, each with probability $\frac{1}{3}$; $y$ may pay out 1 or 2 dollars, each with probability $\frac{1}{2}$, and $z$ pays 2 dollars for sure. Assume that this individual is an expected utility maximizer who prefers more money to less. Which lottery will (s)he choose?

Given that the individual in question “prefers more money to less,” we may assume that her utility (for money) function $u : \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing. Let’s also suppose that $u$ is bounded for good measure – we shall see in Remark 3.1 below that this is without loss of generality. But this is all we know; is this enough information?

\textsuperscript{26}Because of its importance for economics and finance, I will keep coming back to the theory of stochastic dominance throughout the text. For comprehensive accounts, however, you should consult on authorities like Müller and Stoyan (2002) or Shaked and Shanthikumar (2007).
to evaluate these gambles "from the perspective of this individual"? In the case of \(x\) and \(y\) (and \(z\) and \(y\)), the answer is yes. For,

\[
\frac{1}{3}u(1) + \frac{1}{3}u(2) + \frac{1}{3}u(3) \geq \frac{1}{2}u(1) + \frac{1}{2}u(2)
\]

for any self-map \(u\) on \(\mathbb{R}\) with \(u(3) \geq \max\{u(1), u(2)\}\). It follows that \(x \succFSD y\), so our agent is sure to choose \(x\) over \(y\). The upshot is that we are able to reach this conclusion without knowing the exact form of \(u\). This is the power of \(\succFSD\).

And how about \(x\) and \(z\)? Things are a bit more complicated there. For we have neither \(x \succFSD z\) nor \(z \succFSD x\), so we cannot safely predict the behavior of our agent without further information. (Note. It is easily shown that \(E(u(z)) > E(u(x))\) if \(u\) is strictly concave, whereas the opposite inequality holds if \(u\) is strictly convex.)

**Example 3.2.** It is standard practice to model an income distribution as a nonnegative random variable \(x\), where we interpret \(p_x[0, t]\) as the fraction of the population receiving an income of \(t\) dollars or less. (For instance, if \(F_x(1000) = 0.3\), we understand that the set of all individuals each one of whom has at most $1000 constitutes the 30% of the population.) This formulation is quite general. It covers the case of discrete income distributions which is appropriate when one considers a finite population model.\(^{27}\) It also allows one to model the population as consisting of a continuum of individuals (the case where \(F_x\) is strictly increasing over a nondegenerate interval).

For any income distribution \(x\), the per capita income is found as \(E(x)\). On the other hand, given an increasing (social) utility function \(u : \mathbb{R} \to \mathbb{R}\) with respect to which we measure the well-being of an arbitrary member of the society, the aggregate welfare entailed by the income distribution \(x\) may be defined as

\[
W_u(x) := \int_{\mathbb{R}_+} ud\mu_x.
\]

\((W_u\) is sometimes called the utilitarian social welfare function.\(^{28}\)) Of course, there is no obvious way of determining the "right" (social) utility function in assessing total welfare. It is thus of interest to find out exactly when we can make a welfare comparison between two income distributions independently of the choice of the (social) utility function. Indeed, if \(W_u(x) \geq W_u(y)\) for all increasing self-maps \(u\) on \(\mathbb{R}\) (such that \(W_u(x)\) and \(W_u(y)\) are well-defined), which, as we shall see in Remark 3.1, is the same thing as saying that \(x \succFSD y\), then it makes sense to qualify the society with the income distribution \(x\) to be "unambiguously" better off than that with \(y\).

\(^{27}\)A discrete income distribution is a nonnegative vector in a Euclidean space, and can be identified with a simple nonnegative random variable \(x\). For instance, if the society under consideration has \(n\) members and the income level of person \(i\) is \(x_i \geq 0\), then we have \(F_x(t) := \frac{1}{n} |\{i \in [n] : x_i \leq t\}|\) for every real number \(t\). (More on this in Section E.5.2.)

\(^{28}\)Implicit in this is an additive way of aggregating individual utilities. For instance, if the society under consideration consists of \(n\) individuals, then it is easily checked that \(W_u(x) = \frac{1}{n} \sum_{i \in [n]} u(x_i)\) where \(x_i \geq 0\) is the income level of individual \(i\).
These examples point to the importance of the first-order stochastic dominance preorder in economics and finance. But we have a problem here. To be able to say that a random variable \( x \) first-order stochastically dominates another random variable \( y \), we need to verify the inequality \( \mathbb{E}(u \circ x) \geq \mathbb{E}(u \circ y) \) for every increasing and bounded self-map \( u \) on \( \mathbb{R} \), which is a daunting task. Fortunately, there is a nice way to get around this “computational” difficulty.

To get a handle on the problem, let us go in the other direction by first choosing a few interesting \( u \) functions to use in the inequality \( \mathbb{E}(u \circ x) \geq \mathbb{E}(u \circ y) \). For instance, how about trying \( u := 1_{(t, \infty)} \) for any real number \( t \)? This self-map on \( \mathbb{R} \) is increasing and bounded, so if we were to have \( x \preceq_{\text{FSD}} y \), then, necessarily, we would have \( \mathbb{E}(1_{(t, \infty)} \circ x) \geq \mathbb{E}(1_{(t, \infty)} \circ y) \) for every real number \( t \). But, by the Change of Variables Formula, the left-hand side here is \( p_x(t, \infty) \) and the right-hand side is \( p_y(t, \infty) \). We thus found a necessary condition for \( x \preceq_{\text{FSD}} y \) to hold: \( p_x(t, \infty) \geq p_y(t, \infty) \), or equivalently, \( p_x(-\infty, t] \leq p_y(-\infty, t] \) for every real number \( t \). Put differently,

\[
x \preceq_{\text{FSD}} y \implies F_x \leq F_y,
\]

where \( F_x \) is the distribution function of \( x \) and \( F_y \) that of \( y \).

Amazingly, the converse of this observation is also true, that is, the “easily checkable” condition \( F_x \leq F_y \) guarantees \( x \preceq_{\text{FSD}} y \). The proof is not even hard, given what we already know about probability theory. The main difficulty stems from the fact that \( x \) and \( y \) need not be defined on the same probability space. But we can always find two random variables on the probability space \(((0, 1), \mathcal{B}(0, 1), \ell)\) whose distributions equal those of \( x \) and \( y \), respectively. It is these random variables that we will use to make the necessary comparisons. Indeed, we have shown this in Remark C.2.1: If \( z := F_x^{-1} \) and \( w := F_y^{-1} \), where \( F_x^{-1} \) and \( F_y^{-1} \) are the pseudo-inverses of \( F_x \) and \( F_y \), respectively, then \( z \) and \( w \) are (increasing) random variables on \(((0, 1), \mathcal{B}(0, 1), \ell)\) with \( p_x = \ell_z \) and \( p_y = \ell_w \). But, for any \( \omega \in (0, 1) \), we have \( \{ F_x \geq \omega \} \subseteq \{ F_y \geq \omega \} \) (because \( F_x \leq F_y \)), and hence

\[
F_x^{-1}(\omega) = \inf\{ t \in \mathbb{R} : F_x(t) \geq \omega \} \geq \inf\{ t \in \mathbb{R} : F_y(t) \geq \omega \} = F_y^{-1}(\omega).
\]

It follows that \( z \geq w \).\(^{29}\) So, for any increasing self-map \( u \) on \( \mathbb{R} \) such that both \( \mathbb{E}(u \circ x) \) and \( \mathbb{E}(u \circ y) \) exist, the Change of Variables Formula yields

\[
\mathbb{E}(u \circ x) = \int_{\mathbb{R}} u \, dp_x = \int_{\mathbb{R}} u \, d\ell_z = \mathbb{E}(u \circ z) \geq \mathbb{E}(u \circ w) = \mathbb{E}(u \circ y).
\]

In particular, \( x \preceq_{\text{FSD}} y \). We have proved:

\(^{29}\)This is important. I have just shown that, even though \( x \) and \( y \) may be defined on different probability spaces, we are sure to find two random variables \( z \) and \( w \) on a particular probability space, namely \(((0, 1), \mathcal{B}(0, 1), \ell)\), such that (1) \( z \geq w \) (provided \( F_x \leq F_y \)), and (2) probabilistically speaking, \( x \) and \( z \), and \( y \) and \( w \), are indistinguishable from each other. In this sense, then, it makes very good sense to interpret the statement “\( F_x \leq F_y \)” as saying that “\( x \) is stochastically larger than \( y \),” and I’m about to show you that we have \( x \preceq_{\text{FSD}} y \) in this case.
The Fundamental Theorem of Stochastic Dominance 1. (The Univariate Version)
For any random variables \( x \) and \( y \), we have \( x \gtrsim_{\text{FSD}} y \) if, and only if,
\[
p_x(-\infty, t) \leq p_y(-\infty, t) \quad \text{for every } -\infty < t < \infty.
\]

Remark 3.1. Actually, the argument given above establishes a little more than this. Apparently, (23) holds for two random variables \( x \) and \( y \) if the graph of the distribution function of \( x \) lies everywhere under that of \( y \). This means that, for any real number \( t \), the probability that we observe an outcome that yields an \( x \) value less than \( t \) is smaller than the probability of obtaining an outcome (in the experiment that relates to \( y \)) in which \( y \) assumes a value less than \( t \). To say this a little differently, suppose that \( x \) and \( y \) model two gambling situations in the sense that \( x \) stands for the earnings of an individual in the first gamble, and \( y \) for those in the second. Then, \( x \gtrsim_{\text{FSD}} y \) holds if that one will win strictly more than \( $t \) is higher in the first gamble than in the second one, that is, \( p_x(t, \infty) \geq p_y(t, \infty) \), and this, for every \( t \). It is in this sense that the first gamble is unambiguously better than the second one.

The Fundamental Theorem of Stochastic Dominance 1 gives us an alternative way of looking at the preorder \( \gtrsim_{\text{FSD}} \). Apparently, a random variable \( x \) first-order stochastically dominates another random variable \( y \) if the graph of the distribution function of \( x \) lies everywhere under that of \( y \). This means that, for any real number \( t \), the probability that we observe an outcome that yields an \( x \) value less than \( t \) is smaller than the probability of obtaining an outcome (in the experiment that relates to \( y \)) in which \( y \) assumes a value less than \( t \). To say this a little differently, suppose that \( x \) and \( y \) model two gambling situations in the sense that \( x \) stands for the earnings of an individual in the first gamble, and \( y \) for those in the second. Then, \( x \gtrsim_{\text{FSD}} y \) holds if the probability that one will win strictly more than \$t \) is higher in the first gamble than in the second one, that is, \( p_x(t, \infty) \geq p_y(t, \infty) \), and this, for every \( t \). It is in this sense that the first gamble is unambiguously better than the second one.

Example 3.3. [1] For any \( c > 0 \), let \( x_c \) be a random variable that is uniformly distributed on \([0, c]\). Then, for any \( a, b > 0 \), we have \( x_b \gtrsim_{\text{FSD}} x_a \) if \( b \geq a \). Indeed, it is easily checked that \( F_{x_b} \leq F_{x_a} \) if \( a \leq b \), and thus the claim follows from the Fundamental Theorem of Stochastic Dominance 1.

[2] Let \( x \) be a random variable that is uniformly distributed on \([0, 1]\), and \( a \in \mathbb{R} \). Take any random variable \( x_a \) that equals \( a \) with probability one. Then \( x \) and \( x_a \) can be ordered by \( \gtrsim_{\text{FSD}} \) iff \( a \) does not belong to \((0, 1)\). Again, a proof for this is sketched easily by using the Fundamental Theorem of Stochastic Dominance 1.

Example 3.4. For any given positive integer \( m \), take two simple random variables \( x \) and \( y \) whose ranges are contained in \([m]\). Then, by the Fundamental Theorem of Stochastic Dominance 1,
\[
x \gtrsim_{\text{FSD}} y \quad \text{iff} \quad \sum_{i \in [k]} p_x(i) \leq \sum_{i \in [k]} p_y(i) \quad \text{for each } k \in [m-1].
\]
Or, suppose $x$ and $y$ are discrete random variables whose values are bounded from below, and for concreteness, assume that the range of either $x$ or $y$ is countably infinite. Where we enumerate $x(X) \cup y(X)$ as $\{t_1, t_2, \ldots\}$ such that $t_1 < t_2 < \cdots$, we have

$$x \preceq_{FSD} y \iff \sum_{i \in [k]} p_x \{t_i\} \leq \sum_{i \in [k]} p_y \{t_i\} \text{ for each } k \in \mathbb{N}.$$ 

Unfortunately, in many situations (such as when stochastically ranking discrete random variables that are distributed according to a parametric probability distribution), even this characterization is quite difficult to use.\(^{30}\)

**Example 3.5.** [1] Let $x_\lambda$ be an exponentially distributed random variable with parameter $\lambda > 0$. If $0 < \alpha < \beta$, then $1 - e^{-\alpha t} < 1 - e^{-\beta t}$ for all $t > 0$, and it follows from the Fundamental Theorem of Stochastic Dominance 1 that $x_\lambda \succeq_{FSD} x_\beta$. Similarly, it can be checked that if $x_{\mu, \sigma}$ is a normally distributed random variable with parameters $\mu$ and $\sigma^2$, then $x_{\mu, \sigma} \succeq_{FSD} x_{\alpha, \sigma}$ holds for any real numbers $\alpha$ and $\beta$ with $\alpha < \beta$.

[2] Consider an expected utility maximizer (who prefers more money to less). This individual is presented with two portfolios. The payout of the first portfolio is a random variable that is exponentially distributed with parameter $\lambda > 0$. Or, suppose $x$ and $y$ are discrete random variables whose values are bounded from below, and for concreteness, assume that the range of either $x$ or $y$ is countably infinite. Where we enumerate $x(X) \cup y(X)$ as $\{t_1, t_2, \ldots\}$ such that $t_1 < t_2 < \cdots$, we have

$$x \preceq_{FSD} y \iff \sum_{i \in [k]} p_x \{t_i\} \leq \sum_{i \in [k]} p_y \{t_i\} \text{ for each } k \in \mathbb{N}.$$ 

Unfortunately, in many situations (such as when stochastically ranking discrete random variables that are distributed according to a parametric probability distribution), even this characterization is quite difficult to use.\(^{30}\)

Example 3.5. [1] Let $x_\lambda$ be an exponentially distributed random variable with parameter $\lambda > 0$. If $0 < \alpha < \beta$, then $1 - e^{-\alpha t} < 1 - e^{-\beta t}$ for all $t > 0$, and it follows from the Fundamental Theorem of Stochastic Dominance 1 that $x_\lambda \succeq_{FSD} x_\beta$. Similarly, it can be checked that if $x_{\mu, \sigma}$ is a normally distributed random variable with parameters $\mu$ and $\sigma^2$, then $x_{\mu, \sigma} \succeq_{FSD} x_{\alpha, \sigma}$ holds for any real numbers $\alpha$ and $\beta$ with $\alpha < \beta$.

[2] Consider an expected utility maximizer (who prefers more money to less). This individual is presented with two portfolios. The payout of the first portfolio is a random variable that is exponentially distributed with parameter $\frac{1}{2}$, and that of the second is a random variable that is exponentially distributed with parameter $1$. By the previous example, this person prefers the first portfolio. □

In the following set of exercises $x$ and $y$ are two random variables on an arbitrarily given probability space $(X, \Sigma, p)$.

**Exercise 3.1.** For any real number $a$, let $x_a$ be a random variable such that $p_{x_a} \{a\} = 1$.
(a) Prove or disprove: If $p_x \{t, \infty\} = 0$ for some real number $t$, we have $x_a \preceq_{FSD} x$ iff $a \geq \inf \{ t \in \mathbb{R} : p_x \{t, \infty\} = 0 \}$.
(b) If $p_x \{t, \infty\} > 0$ for every real number $t$, can there be a real number $a$ with $x_a \succeq_{FSD} x$?

**Exercise 3.2.** Show that $x \preceq_{FSD} y$ implies $f(x) \preceq_{FSD} f(y)$ for any increasing self-map $f$ on $\mathbb{R}$. Conclude that $x \preceq_{FSD} y$ implies $x^+ \preceq_{FSD} y^+$ and $y^- \preceq_{FSD} x^-$. 

**Exercise 3.3.** Assume that $x$ is a continuous random variable, and take any increasing self-map $f$ on $\mathbb{R}$. Show that $f \circ x \succeq_{FSD} x$ iff $f(t) \geq t$ for each $t \in \mathbb{R}$.

**Exercise 3.4.** Assume that $x$ and $y$ are $\mathbb{Z}_+\text{-valued}$ and $p_x \{k\} > 0$ for each $k \geq 0$. Show that $x \preceq_{FSD} y$ holds if $k \mapsto p_y \{k\} / p_x \{k\}$ is a decreasing map on $\mathbb{Z}_+$, but the converse need not be true.

**Exercise 3.5.** In the statement of the Fundamental Theorem of Stochastic Dominance 1, we could take $\mathcal{U}$ as the class of all increasing and continuous self-maps on $\mathbb{R}$ such that $\mathbb{E}(u \circ x)$ and $\mathbb{E}(u \circ y)$ exist. Prove!

**Exercise 3.6.** (Scarsini-Shaked) Let $x$ and $y$ be nonnegative, and assume that $\mathbb{E}(x) = \mathbb{E}(y) < \infty$. Show that $x \preceq_{FSD} y$ iff $p_x = p_y$.

---

30 To appreciate this, you might want to have a look at the very nice paper by Klemke and Mattner (2010).
Exercise 3.7. For any \( p \in (0, 1) \), let \( x_p \) be a random variable which has the geometric distribution with parameter \( p \). Prove or disprove: \( x_p \gtrsim_{FSD} x_q \) iff \( p \leq q \).

Exercise 3.8. (Klemke-Mattner) For any \( n \in \mathbb{N} \) and \( p \in [0, 1] \), let \( x_{n,p} \) be a random variable which has the binomial distribution with parameters \( n \) and \( p \). Prove: \( x_{n,p} \gtrsim_{FSD} x_{m,q} \) iff \((1-q)^n \geq (1-p)^n \) and \( n \geq m \).

### 3.2 The Multivariate Case

In many applications, the outcome of the experiments of interest are represented by random vectors (instead of random variables), and hence, as we have defined it so far, the first-order stochastic dominance preorder is of no immediate use. It is, however, not difficult to extend the definition of this preorder to the context of random elements of a general (ordered) metric space, thereby enhancing the applicability of stochastic dominance theory. We carry out such an extension in this section.

Given a metric space \( Y \), what we would like to do is to adopt the definition of first-order stochastic dominance to say something like “a \( Y \)-valued random variable \( x \) dominates another such random variable \( y \) if \( \mathbb{E}(u \circ x) \geq \mathbb{E}(u \circ y) \) holds for every increasing and bounded real map on \( Y \).” Good, but to do that we need to have a sense of how to order the members of \( Y \), otherwise the notion of “increasing real map on \( Y \)” would be meaningless. In order to carry out the extension we seek, it appears necessary that we focus on ordered metric spaces (acting as codomains for the random variables under consideration).

**Definition.** A partially ordered metric space is an ordered pair \((Y, \geq)\), where \( Y \) is a metric space and \( \geq \) is a preorder on \( Y \) which is a closed subset of \( Y \times Y \). We say that \((Y, \geq)\) is a partially ordered Polish space if \( Y \) is a complete and separable metric space, and denote this ordered pair simply by \( Y \) (assuming that the partial order \( \geq \) is understood implicitly).

A prominent example of a partially ordered Polish space is, of course, our beloved \( \mathbb{R}^n \) under its usual (componentwise) order. The same goes for \( \mathbb{R}^\infty \) as well. Moreover, for any metric space \( X \), any closed metric subspace of \( B(X) \), in particular \( C_b(X) \), is a partially ordered Polish space under the usual (pointwise) order.

**Definition.** Let \( Y \) be a partially ordered metric space, and \( x \) and \( y \) two \( Y \)-valued random variables. We say that \( x \) first-order stochastically dominates \( y \), and write \( x \gtrsim_{FSD,Y} y \), if

\[
\mathbb{E}(u \circ x) \geq \mathbb{E}(u \circ y)
\]

for every increasing and Borel measurable \( u \in B(Y) \).

When at least one of these inequalities is strict, we write \( x \succ_{FSD,Y} y \).

As every increasing self-map on \( \mathbb{R} \) is Borel measurable (Example C.1.4), this definition reduces to the one we gave in the previous section when \( Y = \mathbb{R} \).

**Example 3.6.** Let \( x \) be an \( \mathbb{R}^2 \)-valued random variable such that \( p_x \{(0, 1)\} = \frac{1}{2} = p_x \{(1, 0)\} \), and for any real number \( a \), let \( x_a \) be an \( \mathbb{R}^2 \)-valued random variable with \( p_{x_a} \{(a, a)\} = 1 \). Then, \( x \gtrsim_{FSD, \mathbb{R}^2} x_a \) iff \( 0 \geq a \), and \( x_a \gtrsim_{FSD, \mathbb{R}^2} x \) iff \( a \geq 1 \). If \( y_a \) is an \( \mathbb{R}^2 \)-valued random variable such that \( p_{y_a} \{(0, 0)\} = \frac{1}{2} = p_{y_a} \{(a, a)\} \), then \( x \) and \( y_a \) cannot be ranked by \( \gtrsim_{FSD, \mathbb{R}^2} \) for any \( a > 0 \). \( \square \)

---

31 By closedness of \( \geq \), here, we mean that, for any two convergent sequences \((\omega_m)\) and \((\nu_m)\) in \( Y \), we have \( \lim \omega_m \geq \lim \nu_m \) whenever \( \omega_m \geq \nu_m \) for each \( m \).
Example 3.7. For any positive integer \( m \), let \( x_m \) be a \( [0, 1] \)-valued random variable such that \( p_{x_m}\{f_m\} = \frac{1}{2} = p_x(\emptyset) \), where \( \emptyset \) is the zero function on \([0, 1]\), and \( f_m(t) := t^m \), \( 0 \leq t \leq 1 \). Then we have \( x_m \preceq_{\text{FSD},C[0,1]} x_{m+1} \) for any \( m = 1, 2, \ldots \). \( \square \)

Exercise 3.9. Let \( Y \) be a partially ordered metric space, and take any two \( Y \)-valued random variables with \( x \preceq_{\text{FSD},Y} y \). Show that, for any increasing map \( f : Y \to \mathbb{R} \), we have \( f \circ x \preceq_{\text{FSD}} f \circ y \). Conversely, suppose \( f \circ x \preceq_{\text{FSD}} f \circ y \) for every increasing \( f : Y \to \mathbb{R} \). Do we necessarily have \( x \preceq_{\text{FSD},Y} y \)?

Exercise 3.10. Let \( x \) be an \( \mathbb{R}^n \)-valued random variable on a probability space \( (X, \Sigma, \mathbf{p}) \). Prove or disprove: \( x + y \succ_{\text{FSD},\mathbb{R}^n} x \) for every \( \mathbb{R}^n \)-valued random variable \( y \) on \( (X, \Sigma, \mathbf{p}) \).

Just as in the univariate case, applying the general first-order stochastic dominance relation requires us to check an integral inequality for every increasing and Borel measurable map in \( \mathcal{B}(Y) \), where \( Y \) is the codomain (of the random variables) of interest. The Fundamental Theorem of Stochastic Dominance 1 proved in the previous section gives us a nice way to deal with this difficulty in the case of real-valued random variables \( x \) and \( y \) by showing that the statement “\( x \preceq_{\text{FSD}} y \)” is equivalent to the statement “\( p_x(\omega, t) \leq p_y(\omega, t) \)” for every real number \( t \)” We now wish to prove a similar characterization in the general case. But for this, we need to be able to find the counterpart of the interval \( (-\infty, t] \) within an arbitrary partially ordered metric space. Well, here is an observation: In \( \mathbb{R} \), a closed set \( S \) has the property that any real number smaller than an element of \( S \) also belongs to \( S \) if, and only if, \( S = (-\infty, t] \) for some \( t \in \mathbb{R} \). This is nice because it gives us a way of capturing the subsets of a partially ordered metric space that behave “like” the intervals \( (-\infty, t] \).

To pursue this idea, let us agree to say that a subset \( S \) of a partially ordered metric space \( Y \) is decreasing if, for every \( (\omega, \nu) \in S \times Y \) with \( \omega \geq \nu \), we have \( \nu \in S \). Then, a natural conjecture is that the statement “\( x \preceq_{\text{FSD}} y \)” is equivalent to the statement “\( p_x(S) \leq p_y(S) \)” for all decreasing closed subset \( S \) of \( Y \)” The main result of this section shows that this conjecture is true, at least when \( Y \) is well-behaved.\(^{32}\)

The Fundamental Theorem of Stochastic Dominance 1. (The General Version) Let \( Y \) be a partially ordered Polish space, and \( x \) and \( y \) two \( Y \)-valued random variables. Then, \( x \preceq_{\text{FSD},Y} y \) holds if, and only if,

\[
p_x(T) \leq p_y(T) \quad \text{for every decreasing closed subset } T \text{ of } Y. \tag{24}
\]

Proof. Assume that \( x \preceq_{\text{FSD},Y} y \), and take any decreasing \( T \in \mathcal{C}_Y \). Then, \( -1_T \) is an increasing and Borel measurable map in \( \mathcal{B}(Y) \), so we have

\[
-p_x(T) = -E(1_T \circ x) = E(-1_T \circ x) \geq E(-1_T \circ y) = -p_y(T)
\]

by the Change of Variables Formula. That is, \( p_x(T) \leq p_y(T) \).

Conversely, assume that (24) holds. We first wish to improve (24) to the following:

\[
p_x(S) \leq p_y(S) \quad \text{for every decreasing Borel subset } S \text{ of } Y. \tag{25}
\]

So, take any decreasing Borel subset \( S \) of \( Y \), and fix an arbitrary \( \varepsilon > 0 \). By tightness of \( p_x \) (Corollary B.2.8), we can find a compact set \( K \) in \( X \) such that \( K \subseteq S \) and \( p_x(S) - p_x(K) < \varepsilon \). (It is at this point that we make use of the fact that \( Y \) is Polish.) Now define the set

\[
T := \{ \nu \in Y : \omega \geq \nu \text{ for some } \omega \in K \}.
\]

\(^{32}\)This result was put on record first by Kamae, Krengel and O’Brien (1977).
Clearly, $T$ is a decreasing subset of $Y$ and $K \subseteq T \subseteq S$. We claim that $T$ is closed. To see this, take any sequence $(\nu_m)$ in $T$ that converges to some element $\nu$ of $Y$. Then, by definition of $T$, there is a sequence $(\omega_m)$ in $K$ such that $\omega_m \geq \nu_m$ for each $m$. Since $K$ is compact, there is a subsequence $(\omega_{m_k})$ of $(\omega_m)$ which converges in $K$. As $\geq$ is a closed subset of $Y \times Y$, we have $\lim \omega_{m_k} \geq \lim \nu_{m_k} = \nu$, and it follows that $\nu \in T$. Conclusion: $T$ is closed. Since $K \subseteq T \subseteq S$ (and hence $p_x(K) \leq p_x(T) \leq p_x(S)$), we may thus apply (24) to find
\[
p_x(S) \leq p_x(T) + \varepsilon \leq p_y(T) + \varepsilon \leq p_y(S) + \varepsilon.
\]
Since $\varepsilon > 0$ is arbitrary here, there follows (25).

Now take any increasing and Borel measurable map $u \in B(Y)$, and note that $\{u \leq a\}$ is a decreasing Borel subset of $Y$ for any $a > 0$.\footnote{This set need not be closed, which is why I went through the intermediate step of extending (24) to (25).} Put $a := \sup u(Y)$, and, for any positive integer $m$, define
\[
u_m := a - \frac{1}{m} \sum_{i=0}^{\infty} 1 \{u \leq \alpha - \frac{i}{m}\}.
\]
Since $u$ is bounded, we have $\{u \leq a\} = \emptyset$ for $a$ small enough, so, for every positive integer $m$, there is a $k_m \in \mathbb{N}$ such that
\[
u_m = a - \frac{1}{m} \sum_{i=0}^{k_m} 1 \{u \leq \alpha - \frac{i}{m}\}.
\]
(In particular, this shows that each $\nu_m$ is well-defined.) Then, by (25),
\[
\int_Y \nu_m d\nu_x = a - \frac{1}{m} \sum_{i=0}^{k_m} \nu_x \{u \leq \alpha - \frac{i}{m}\} \geq a - \frac{1}{m} \sum_{i=0}^{k_m} \nu_y \{u \leq \alpha - \frac{i}{m}\} = \int_Y \nu_m d\nu_y
\]
for each $m$. But it is easy to check that $\nu_m \rightarrow u$ uniformly, so applying Exercise 2.9 (or Observation 1.1) and the Change of Variables Formula yields $E(u \circ x) \geq E(u \circ y)$, as we sought.

**Remark 3.2.** In Section 3.1, we were able to prove the following fact: For any random variables $x$ and $y$ with $x \geq_{\text{FSD}} y$, there exist random variables $z$ and $w$ on a common probability space such that $p_x = p_z$, $p_y = p_w$ and $z \geq w$. A famous theorem of Strassen (1965) says that the following extension of this result is true: For any partially ordered Polish space $Y$, and any two $Y$-valued random variables $x$ and $y$ with $x \geq_{\text{FSD},} Y$, there exist $Y$-valued random variables $z$ and $w$ on a common probability space such that $p_x = p_z$, $p_y = p_w$ and $z \geq_{\text{a.s.}} w$. The proof of this result is, however, beyond the scope of the present text.

## 4 Elementary Probability Inequalities

### 4.1 Jensen’s Inequality

There are numerous inequalities in probability theory which allows us to have various estimates about the structure of a random variable defined on a probability space in terms of the expectation of that random variable. An especially useful inequality in this regard says that, under fairly general conditions, the expectation of a concave transformation of a random variable $x$ is always less than the same transformation
of the expectation of \( x \). This inequality is widely used in economics and information theory. It goes by the name of *Jensen’s Inequality*, for a preliminary version of it was first proved by Johan Jensen in 1906.

**Jensen’s Inequality.** Let \( x \) be an integrable random variable on a probability space \((X, \Sigma, p)\), \( I \) an open interval that contains \( x(X) \), and \( \varphi : I \to \mathbb{R} \) a concave function. Then, \( \varphi \circ x \in \mathcal{L}^0(X, \Sigma) \) and

\[
\mathbb{E}(\varphi \circ x) \leq \varphi(\mathbb{E}(x)).
\]

If \( \varphi \) is convex, then the inequality goes the other direction.

**Proof.** Since \( \varphi \in \mathbb{R}^I \) is concave, it is continuous on \( I \) (Corollary 2.7 of Appendix 1), and hence \( \varphi \in \mathcal{L}^0(I, \mathcal{B}(I)) \). It follows that \( \varphi \circ x \) is a random variable on \((X, \Sigma)\). To prove the main part of our proposition, we use Corollary 2.9 of Appendix 1 to find two real numbers \( a \) and \( b \) such that \( a\mathbb{E}(x) + b = \varphi(\mathbb{E}(x)) \) and \( \varphi(s) \leq as + b \) for all \( s \in I \). (See Figure D.2.) Then \( \varphi(x(\omega)) \leq ax(\omega) + b \) for all \( \omega \in X \), which implies that \( \varphi \circ x \) is integrable, and

\[
\mathbb{E}(\varphi \circ x) \leq \mathbb{E}(ax + b) = a\mathbb{E}(x) + b = \varphi(\mathbb{E}(x))
\]

as we sought.

**FIGURE D.2 ABOUT HERE**

**Warning.** Jensen’s Inequality fails in a major way in the context of infinite measure spaces. For instance, consider the measure space \((\mathbb{N}, 2^\mathbb{N}, \mu)\) where \( \mu \) is the counting measure on \( 2^\mathbb{N} \), and put \( I := \mathbb{R}_{++} \). Then, where \( x : X \to \mathbb{R} \) and \( \varphi : I \to \mathbb{R} \) are defined by \( x(i) := 1/i^2 \) and \( \varphi(t) := -t^2 \), respectively, we have \( \int_X (\varphi \circ x) d\mu = -\pi^4/90 > -\pi^2/36 = \varphi\left(\int_X x d\mu\right) \). However where \((X, \Sigma, \mu)\) is any finite measure space, we have

\[
\frac{1}{\mu(X)} \int_X (\varphi \circ x) d\mu \leq \varphi\left(\frac{1}{\mu(X)} \int_X x d\mu\right)
\]

where \( x, I \) and \( \varphi \) are as in Jensen’s Inequality.

**Exercise 4.1.** Give an example to show that the hypothesis \( \mathbb{E}(x) < \infty \) cannot be omitted in Jensen’s Inequality.

We will have plenty of occasions later on to make use of this celebrated inequality. For your immediate enjoyment, here are some quick applications.

In the following set of exercises \( x \) and \( y \) stand for two integrable random variables on a given probability space \((X, \Sigma, p)\).
Exercise 4.2. (The Arithmetic-Geometric Mean Inequality) Take any positive integer \( m \), and real numbers \( a_1, \ldots, a_m \geq 0 \) and \( r_1, \ldots, r_m > 0 \) such that \( \sum_{i=1}^m 1/r_i = 1 \). Use the fact that \( t \mapsto \log t \) is a concave function on \( \mathbb{R}_+^+ \) to prove that
\[
\sum_{i \in [m]} \frac{a_i}{r_i} \geq \prod_{i \in [m]} a_i^{1/r_i}
\]
with equality iff \( a_1 = \cdots = a_m \).\(^{34}\)

Exercise 4.3. (Ordering of Power Means) Let \( n \) be a positive integer, and \( \theta \) a nonzero real number. The \( \theta \)-power mean is the real function \( M_\theta \) on \( \mathbb{R}_+^n \), defined by
\[
M_\theta(a_1, \ldots, a_n) := \left( \frac{1}{n} \sum_{i \in [n]} a_i^\theta \right)^{1/\theta}.
\]
Show that \( M_\alpha \leq M_\beta \) for any nonnegative real numbers \( \alpha \) and \( \beta \) with \( \alpha < \beta \).

Exercise 4.4. Suppose that \( x \) and \( y \) are comonotonic, that is,
\[
(x(\omega) - x(\nu))(y(\omega) - y(\nu)) \geq 0 \quad \text{for every } \omega, \nu \in X.
\]
Prove that \( \mathbb{E}(xy) \geq \mathbb{E}(x)\mathbb{E}(y) \). Thus: \( \mathbb{E}(x)\mathbb{E}(\frac{1}{x}) \leq 1 \) whenever \( x > 0 \).

Exercise 4.5. \(^{11}\) (The Mean-Median Inequality) A median of \( x \) is defined to be any number \( m \) such that \( p_x(1, m] \geq \frac{1}{2} \) and \( p_x(m, \infty) \geq \frac{1}{2} \). We denote an arbitrarily fixed median of \( x \) as \( \text{med}(x) \).
(a) Show that \( \text{med}(x) \in \arg \min \{ \mathbb{E}(|x - a|) : a \in \mathbb{R} \} \).
(b) Use Jensen’s Inequality to prove that
\[
|\mathbb{E}(x) - \text{med}(x)| \leq \sqrt{\mathbb{E}((x - \mathbb{E}(x))^2)}.
\]

Insight. The median and the expectation of a random variable (with finite mean) cannot differ more than one standard deviation.

Exercise 4.6. (Converse of Jensen’s Inequality) Take any \( z \in \mathcal{L}^0(\mathcal{B}(\mathbb{R})) \) with \( z(\mathbb{R}) = [0, 1] \), and let \( \varphi \in \mathcal{C}[0, 1] \). Prove: If \( \mathbb{E}_p(\varphi(z)) \leq \varphi(\mathbb{E}_p(z)) \) for every \( p \in \Delta(\mathbb{R}) \), then \( \varphi \) must be concave on \( [0, 1] \).

The following exercise provides a Jensen-like inequality for functions that are not necessarily concave.

Exercise 4.7. \(^{11}\) (Abramovich-Jameson-Sinnamon) A function \( \varphi \in \mathbb{R}^{[0, \infty)} \) is said to be subquadratic if, there exists a function \( C \in \mathbb{R}^{[0, \infty)} \) such that
\[
\varphi(b) - \varphi(a) - \varphi(b - a) \leq C(a)(b - a)
\]
for every \( a, b \geq 0 \).
(a) Show that if \( \varphi \) is a nonpositive subquadratic function, then it is concave.
(b) Show that a subquadratic function need not be concave.

\(^{34}\)I assume in this exercise that you are familiar with the basic properties of the exponential and logarithmic functions. (As an application of our work on integration theory, these properties will be derived from scratch in Section E.1.6.)
(c) Assume that \( x \) is nonnegative, and take any subquadratic \( \varphi \in \mathcal{C}(\mathbb{R}_+) \). Prove: \( \varphi \circ x \in L^0(X, \Sigma) \) and
\[
\mathbb{E}(\varphi(x) - \varphi(|x - \mathbb{E}(x)|)) \leq \varphi(\mathbb{E}(x)).
\]
This inequality holds as an equality if \( \varphi(t) = -t^2 \) for all \( t \geq 0 \).

*Exercise 4.8.* (A Generalization of Jensen’s Inequality) Let \( n \) be a positive integer and take any integrable random variables \( x_1, \ldots, x_n \) on \((X, \Sigma, p)\). Let \( S \) be a convex subset of \( \mathbb{R}^n \) that contains \( x_1(X) \times \cdots \times x_n(X) \), and \( \varphi : S \to \mathbb{R} \) a concave function such that \( \varphi(x_1, \ldots, x_n) \in L^0(X, \Sigma) \). (If \( S \) is open, then concavity of \( \varphi \) implies its continuity, so the measurability requirement becomes redundant.) Prove that
\[
\mathbb{E}(\varphi(x_1, \ldots, x_n)) \leq \varphi(\mathbb{E}(x_1), \ldots, \mathbb{E}(x_n)). \quad \tag{35}
\]

### 4.2 The Chebyshev-Bienaymé Inequality

Another fundamental inequality of probability theory is the Chebyshev-Bienaymé Inequality, which is often invoked to derive various forms of the laws of large numbers.\(^{36}\)

This inequality is an easy consequence of the following simple result which can also be used to obtain other sorts of estimations for tail probabilities.

**Lemma 4.1.** Let \( Y \) be a metric space, and \( x \) a \( Y \)-valued random variable on a probability space \((X, \Sigma, p)\). Then, for any continuous \( \varphi : Y \to \mathbb{R}_+ \) and real number \( \lambda > 0 \), we have
\[
p \{ \varphi \circ x \geq \lambda \} \leq \frac{1}{\lambda} \mathbb{E}(\varphi \circ x).
\]

**Proof.** It is plain that \( \varphi \circ x \in L^0(X, \Sigma) \). Moreover, where \( S := \{ \varphi \circ x \geq \lambda \} \), we have \( \varphi \circ x \geq \lambda 1_S \), so, by the monotonicity of the Lebesgue integral, \( \mathbb{E}(\varphi \circ x) \geq \mathbb{E}(\lambda 1_S) = \lambda p(S) \), and we are done. \( \blacksquare \)

\(^{35}\)[Only for those who have studied Section 2.5] There are various ways in which we can extend this fact to the context of Bochner integrable random variables. Here is an example: Let \((Y, \geq)\) be an ordered linear space (Section 1.4 of Appendix 3), and suppose that \( Y \) is a Banach space such that \( \geq \) is closed in \( Y \times Y \). Take any \( x \in BC_{1,Y}(X, \Sigma, p) \), and let \( S \) be a closed and convex subset of \( Y \) with \( x(X) \subseteq S \). Then, \( \mathbb{E}(x) \in S \) by Proposition 2.10. Now suppose \( \varphi : S \to Y \) is a continuous function such that \( \varphi \circ x \) is Bochner integrable and \( \varphi(\frac{1}{2} \nu_1 + \frac{1}{2} \nu_2) \geq \frac{1}{2} \varphi(\nu_1) + \frac{1}{2} \varphi(\nu_2) \) for any \( \nu_1, \nu_2 \in Y \). Then, \( \varphi(\mathbb{E}(x)) \geq \mathbb{E}(\varphi \circ x) \). This is not particularly difficult to prove, but it does require the use of a general separating hyperplane argument. For a complete proof, and several other extensions of Jensen’s inequality for the Bochner integral, see Perlman (1974).

\(^{36}\)This inequality is often referred to as Chebyshev’s Inequality, but only a slightly weaker version of it was, in fact, proved by Irénée-Jules Bienaymé in 1853. The version that we consider here was given by Pafnuty Chebyshev in 1867 in order to derive a version of the so-called Weak Law of Large Numbers. (This result is considered the first limit theorem of probability theory.) But Chebyshev’s work goes well beyond probability theory. Indeed, he made significant contributions to number theory – for instance, the fact that there is a prime number between \( n \) and \( 2n \), for any integer \( n \geq 3 \), is proved by him (although this was first conjectured by François Bertrand). Perhaps more importantly, he was one of the major figures responsible from the development of approximation theory in the second half of the 19th century.
**Warning.** Nonnegativity of $\varphi$ is essential for the validity of Lemma 4.1. For instance, if $x$ is a random variable that takes value 1 with probability $\frac{1}{2}$ and $-1$ with probability $\frac{1}{2}$, then $\mathbb{E}(x) = 0$, while, for any $0 < \lambda < 1$, the probability of the event $\{x \geq \lambda\}$ is $\frac{1}{2}$.

Applying Lemma 4.1 with $Y$ being a normed linear space, and $\varphi$ being the norm $\| \cdot \|_Y$ yields:

**Markov’s Inequality.** Let $Y$ be a normed linear space, and $x$ a $Y$-valued random variable on a probability space $(X, \Sigma, \mathbf{P})$. Then, for any real number $\lambda > 0$, we have

$$\mathbf{P} \left\{ \|x\|_Y \geq \lambda \right\} \leq \frac{1}{\lambda^2} \mathbb{E}(\|x\|_Y).$$

In particular,

$$\mathbf{P} \left\{ |x| \geq \lambda \right\} \leq \frac{1}{\lambda^2} \mathbb{E}(|x|) \quad \text{for any random variable } x \text{ on } (X, \Sigma, \mathbf{P}).$$

Applying Lemma 4.1 with $\varphi$ being the map $t \mapsto t^2$ on $\mathbb{R}$ yields:

**The Chebyshev-Bienaymé Inequality.** For any random variable $x$ defined on a probability space $(X, \Sigma, \mathbf{P})$, and any real number $\lambda > 0$, we have

$$\mathbf{P} \left\{ |x| \geq \lambda \right\} \leq \frac{1}{\lambda^2} \mathbb{E}(x^2).$$

In the following set of exercises $x$ stands for a random variable on a given probability space $(X, \Sigma, \mathbf{P})$.

**Exercise 4.9.** Prove: For any increasing self-map $\psi$ on $\mathbb{R}_+$ and real number $\lambda > 0$, we have

$$\mathbf{P} \left\{ |x| \geq \lambda \right\} \leq \frac{\mathbb{E}(\psi(|x|))}{\psi(\lambda)}.$$

In particular:

$$\mathbf{P} \left\{ |x| \geq \lambda \right\} \leq e^{-\lambda^2} \mathbb{E}(e^{|x|}).$$

**Exercise 4.10.** When $x$ is integrable, the Chebyshev-Bienaymé Inequality is often stated as

$$\mathbf{P} \left\{ |x - \mathbb{E}(x)| \geq \lambda \right\} \leq \frac{1}{\lambda^2} \mathbb{V}(x), \quad \lambda > 0.$$

Prove this.

**Exercise 4.11.** The Chebyshev-Bienaymé Inequality may hold as an equality. Verify this by using a random variable $x$ on $((0, 1), \mathcal{B}(0, 1))$ whose distribution function is $\frac{1}{2} \mathbf{1}_{(-\alpha, a)} + \frac{1}{2} \mathbf{1}_{(a, \infty)}$ for some $a > 0$. (How do we know that there is such a random variable?)

**Exercise 4.12.** Assume that both $\mathbb{E}(x)$ and $\mathbb{V}(x)$ are finite.

(a) Show that

$$\mathbf{P} \left\{ x - \mathbb{E}(x) \geq \alpha \right\} \leq \mathbf{P} \left\{ (x - \mathbb{E}(x) + \beta)^2 \geq (\alpha + \beta)^2 \right\} \quad \text{for every } \alpha, \beta > 0.$$

(b) *(Cantelli’s Inequality)* Minimize the right-hand side of the above inequality over $\beta > 0$, and then use Markov’s Inequality, to find

$$\mathbf{P} \left\{ x - \mathbb{E}(x) \geq \alpha \right\} \leq \frac{\mathbb{V}(x)}{\mathbb{V}(x) + \alpha^2} \quad \text{for any } \alpha > 0.$$
5 Spaces of Integrable Random Variables

5.1 \( L^p \) Spaces

In this section we introduce some classical Banach spaces which are defined through Lebesgue integration. We start our discussion in the general context of arbitrary measure spaces, but given our present objectives, we will shortly confine our analysis to the case of probability spaces.

First Impressions

Let \((X, \Sigma, \mu)\) be a measure space. We have earlier considered integrable members of \( L^0(X, \Sigma) \) – these are the measurable real maps \( f \) on \((X, \Sigma)\) such that \( \int_X |f| \, d\mu < \infty \). The collection of all such functions is denoted as \( L^1(X, \Sigma, \mu) \). More generally, for any real number \( p \geq 1 \), we define \( L^p(X, \Sigma, \mu) \) as the set of all random variables \( f \) on \((X, \Sigma)\) such that \( |f|^p \) is integrable. In other words,

\[
L^p(X, \Sigma, \mu) := \left\{ f \in L^0(X, \Sigma) : \int_X |f|^p \, d\mu < \infty \right\}.
\]

Any one member of \( L^p(X, \Sigma, \mu) \) is said to be \( p \)-integrable, but as you know, we refer to 1-integrable random variables simply as integrable. It is also quite common to refer to 2-integrable random variables as square-integrable.

It is plain that \( L^p(X, \Sigma, \mu) \) is closed under (pointwise) scalar multiplication. It is also closed under pointwise addition. Indeed, if \( f, g \in L^p(X, \Sigma, \mu) \), then, because

\[
|f + g|^p \leq (|f| + |g|)^p \leq (2 \max\{|f|, |g|\})^p \leq 2^p(|f|^p + |g|^p),
\]

we find that \( |f + g|^p \) is integrable, that is, \( f + g \in L^p(X, \Sigma, \mu) \). It follows that \( L^p(X, \Sigma, \mu) \) is a linear space (under the usual operations of addition and scalar multiplication). Moreover, the Lebesgue integral is a linear functional on this space.

There is a natural way of making \( L^p(X, \Sigma, \mu) \) a seminormed linear space.\(^{37}\) To this end, we define the real map \( \|\cdot\|_p \) on \( L^p(X, \Sigma, \mu) \) by

\[
\|f\|_p := \left( \int_X |f|^p \, d\mu \right)^{1/p}.
\]

It is plain that this map is absolutely homogeneous and it maps the zero function on \( X \) to 0. It is also subadditive, and hence it is a seminorm on \( L^p(X, \Sigma, \mu) \), but this is hardly a trivial observation. The following exercises aim to prepare you for the proof of this result.

\(^{37}\)Throughout this section (and the next) I assume familiarity with the basics of normed linear spaces. You may want to quickly review Sections 1.1-2, 2.1-2 and 2.4 of Appendix 3 before reading the present discussion.
In the following set of exercises \((X, \Sigma, \mu)\) stands for an arbitrarily fixed measure space, and \(p, q \in [1, \infty)\).

**Exercise 5.1.** Assume \(p > 1\) and \(\frac{1}{p} + \frac{1}{q} = 1\).

(a) (Young’s Inequality) Use the Arithmetic-Geometric Mean Inequality (Exercise 4.2) to show that

\[
ab \leq \frac{a^p}{p} + \frac{b^q}{q}
\]

for every \(a, b \geq 0\).

(b) (Hölder’s Inequality) Take any \(f \in L^p(X, \Sigma, \mu)\) and \(g \in L^q(X, \Sigma, \mu)\). Assume that \(\|f\|_p > 0\) and \(\|g\|_q > 0\) (otherwise what we are after is trivially true), use Young’s Inequality with \(a = \frac{|f(\omega)|}{\|f\|_p}\) and \(b = \frac{|g(\omega)|}{\|g\|_q}\) for any \(\omega \in X\), and integrate the resulting inequality to find

\[
\int_X |fg| d\mu \leq \|f\|_p \|g\|_q.
\]

The following special case of Hölder’s Inequality is worth mentioning:

**The Cauchy-Schwarz Inequality.** \(\|fg\|_1 \leq \|f\|_2 \|g\|_2\) for every \(f, g \in L^2(X, \Sigma, \mu)\).

**Exercise 5.2.** (Ordering of \(L^p\)-norms) Assume that \(\mu\) is a finite measure and \(p \leq q\). Use Hölder’s Inequality to show that

\[
\|f\|_p \leq \|f\|_q
\]

for any \(f \in L^q(X, \Sigma, \mu)\), and conclude that

\[
L^p(X, \Sigma, \mu) \supseteq L^q(X, \Sigma, \mu).
\]

(Compare this result with the observation noted in Exercise 4.3.)

**Exercise 5.3.** (Generalization of Exercise 5.2) Assume that \(p \leq q\) and take any \(f \in L^q(X, \Sigma, \mu)\) such that \(\mu(S(f)) < \infty\), where \(S(f) := \{\omega \in X : f(\omega) \neq 0\}\). Use Hölder’s Inequality to show that

\[
\|f\|_p \leq \mu(S(f))^{\frac{1}{p} - \frac{1}{q}} \|f\|_q.
\]

**Exercise 5.4.** Prove: If \(1 \leq p \leq q < \infty\), then \(L^q(X, \Sigma, p) \supseteq L^p(X, \Sigma, p)\) if and only if there is no sequence \((A_m)\) in \(\Sigma\) such that \(p(A_m) > 0\) for each \(m\) and \(p(A_m) \to 0\).

**Exercise 5.5.** Assume \(p > 1\) and \(\frac{1}{p} + \frac{1}{q} = 1\). Assume that \(\mu\) is a probability measure, and use Hölder’s Inequality to prove the following: If \((x_m)\) and \((y_m)\) are two sequences in \(L^1(X, \Sigma, \mu)\) such that \(|x_m|^p\) and \(|y_m|^q\) are uniformly integrable, then \((x_my_m)\) is uniformly integrable.

We are now prepared for:

**Minkowski’s Inequality.** For any \(p \geq 1\) and \(f, g \in L^p(X, \Sigma, \mu)\),

\[
\|f + g\|_p \leq \|f\|_p + \|g\|_p.
\]

**Proof.** For \(p = 1\), the claim is immediate from the Triangle Inequality (for the absolute value function) and the monotonicity of the Lebesgue Integral. We thus assume that \(p > 1\), and let \(q := \frac{p}{p-1}\). Then, again by the Triangle Inequality,

\[
\|f + g\|_p^p \leq \int_X |f + g|^{p-1} |f| d\mu + \int_X |f + g|^{p-1} |g| d\mu.
\]
But by Hölder’s Inequality,
\[ \int_X |f + g|^{p-1} |f| \, d\mu \leq \left( \int_X |f + g|^{(p-1)} \right)^{\frac{1}{p}} \left( \int_X |f|^p \right)^{\frac{1}{q}} = \|f + g\|_{p/q} \|f\|_p, \]
and similarly, \( \int_X |f + g|^{p-1} |g| \, d\mu \leq \|f + g\|_{p/q} \|g\|_p. \) It follows that
\[ \|f + g\|_p \leq \|f + g\|_{p/q} \left( \|f\|_p + \|g\|_p \right), \]
and our assertion becomes a consequence of the fact that \( p - \frac{2}{q} = 1. \)

In view of Minkowski’s Inequality, we now see that \( (L^p(X, \Sigma, \mu), \| \cdot \|_p) \) is a semi-normed linear space. It is not a normed linear space proper, because \( \| \cdot \|_p \) identifies any two random variables that are distinct from each other only on a negligible set with respect to \( \mu. \) In other words, the map \( (f, g) \mapsto \|f - g\|_p \) is a semimetric, but it is not a metric, for it fails to separate points in \( L^p(X, \Sigma, \mu). \) The problem is that \( \|f\|_p = 0 \) does not yield \( f = 0, \) it implies only that \( f =_{a.s.} 0. \) But this is only a technical wrinkle, because we have \( \|f\|_p = 0 \) iff \( f =_{a.s.} 0. \) Therefore, following the idea sketched in Remark 1.1 of Appendix 2, we can think of \( L^p(X, \Sigma, \mu) \) as a normed (and hence metric) space (relative to the map \( \| \cdot \|_p \)) by identifying those random variables that are equal to each other almost surely.

To say all this a bit more formally, let us define the equivalence relation \( \approx \) on \( L^p(X, \Sigma, \mu) \) by \( f \approx g \) iff \( f =_{a.s.} g, \) and make the quotient set \( L^p(X, \Sigma, \mu) / \approx \) a linear space by defining \( [af] \approx [g] \) and \( [f] \approx [g] \approx [f + g] \) for all \( f, g \in L^p(X, \Sigma, \mu) \) and \( a \in \mathbb{R}. \) Next, define the norm \( \| \cdot \|'_p \) on \( L^p(X, \Sigma, \mu) / \approx \) by \( \|[f] \|'_p := \|f\|_p. \) It is easily checked that \( (L^p(X, \Sigma, \mu) / \approx, \| \cdot \|'_p) \) is a normed linear space, and the metric induced by \( \| \cdot \|_p \) says that the “distance” between \( [f] \) and \( [g] \) is \( \|f - g\|_p \) for any \( p \)-integrable \( f, g \in L^p(X, \Sigma). \) When we refer to a metric property of \( L^p(X, \Sigma, \mu) \) (such as completeness), we mean that property for the metric space \( L^p(X, \Sigma, \mu) / \approx. \) For instance, when we say that \( L^p(X, \Sigma, \mu) \) is a Banach space, what we really mean is that \( L^p(X, \Sigma, \mu) / \approx \) is Banach space (relative to \( \| \cdot \|'_p \)). So long as we keep this in mind, there is really no problem in treating \( L^p(X, \Sigma, \mu) \) “as if” it is a normed (and hence metric) linear space, and we will follow this practice here on out.

**Remark 5.1.** Viewed as normed linear spaces, \( L^p \) spaces generalize several other spaces of interest. For instance, if \( \mu \) is the counting measure on \( 2^{[n]} \), then \( L^2([n], 2^{[n]}, \mu) \) is none other than the \( n \)-dimensional Euclidean space (while, for any \( p \geq 1, L^p([n], 2^{[n]}, \mu) \) is the same as \( \mathbb{R}^n \) with the \( p \)-norm). Similarly, where \( p \in [1, \infty) \) and \( \mu \) is the counting measure on \( 2^\mathbb{N} \), then \( L^p(\mathbb{N}, 2^\mathbb{N}, \mu) \) is none other than the familiar normed linear space \( \ell^p. \) (See Example 1.5 of Appendix 2.)

**Completeness of \( L^p \) Spaces**

65
A very important property of $\mathcal{L}^p$ spaces is their completeness. That is, viewed as a normed linear space, any such space is a Banach space. We prove this next in the context of probability spaces.

**The Riesz-Fischer Theorem.** Let $(X, \Sigma, p)$ be a probability space and $1 \leq p < \infty$. Then $\mathcal{L}^p(X, \Sigma, p)$ is a Banach space.

**Proof.** By Proposition 2.19 of Appendix 3, it is enough to show that every absolutely convergent series in $\mathcal{L}^p(X, \Sigma, p)$ is convergent. That is, given any sequence $(x_m)$ in $\mathcal{L}^p(X, \Sigma, p)$ with $\sum_{i=0}^{\infty} \|x_i\|_p < \infty$, it suffices to show that there is some $x \in \mathcal{L}^p(X, \Sigma, p)$ such that $\left\| \sum_{i=0}^{m} x_i - x \right\|_p \rightarrow 0$. The natural candidate for $x$ is, of course, $\sum_{i=0}^{\infty} x_i$, that is, the pointwise limit of the sequence $(\sum_{i=0}^{m} x_i)$.

But we should be careful here, as this map need not be real-valued. Fortunately, we can easily show that it is almost surely real-valued. To this end, define $y := \sum_{i=0}^{\infty} |x_i|$, which is clearly a $[0, \infty]$-valued random variable on $(X, \Sigma)$. By the Monotone Convergence Theorem 1, continuity of the maps $t \mapsto |t|^p$ and $t \mapsto t^{1/p}$ on $\mathbb{R}_+$, and Minkowski’s Inequality,

$$\|y\|_p = \left( \lim_{m \to \infty} \int_X \left( \sum_{i=0}^{m} |x_i| \right)^p \, dp \right)^{1/p} = \lim_{m \to \infty} \left\| \sum_{i=0}^{m} |x_i| \right\|_p \leq \lim_{m \to \infty} \sum_{i=0}^{\infty} \|x_i\|_p,$$

whence $\|y\|_p \leq \sum_{i=0}^{\infty} \|x_i\|_p$. We conclude that $\mathbb{E}(y^p)$ is a real number. Then, obviously, $p\{y^p = \infty\} = 0$ must be true. (Yes?) This implies $p\{y = \infty\} = 0$, as we sought.

A good candidate for the limit of the sequence $(\sum_{i=0}^{m} x_i)$ (with respect to $\| \cdot \|_p$) thus equals the pointwise limit of the sequence $\{\sum_{i=0}^{\infty} x_i\}$ on the event $\{\sum_{i=0}^{\infty} x_i \in \mathbb{R}\}$, but it can be defined arbitrarily on the complement of that event. (After all, $\| \cdot \|_p$ assigns the same number to almost surely equal members of $\mathcal{L}^p(X, \Sigma, p)$.) Let us then consider the map $x : X \rightarrow \mathbb{R}$ with

$$x(\omega) := \begin{cases} \lim_{m \to \infty} \sum_{i=0}^{m} x_i(\omega), & \text{if } y(\omega) \in \mathbb{R} \\ 0, & \text{otherwise}. \end{cases}$$

We know that $x \in \mathcal{L}^0(X, \Sigma)$. (Yes?) Moreover, $|x| \leq y$, so we have $\mathbb{E}(|x|^p) \leq \mathbb{E}(y^p) < \infty$. Conclusion: $x \in \mathcal{L}^p(X, \Sigma, p)$. It remains to prove that $\left\| \sum_{i=0}^{m} x_i - x \right\|_p \rightarrow 0$. But, again by using the Monotone Convergence Theorem 1, continuity of the maps $t \mapsto |t|^p$ and $t \mapsto t^{1/p}$ on $\mathbb{R}_+$, and Minkowski’s Inequality, we find

$$\left\| x - \sum_{i=0}^{m} x_i \right\|_p = \left( \int_X \left( \sum_{i=m+1}^{\infty} x_i \right)^p \, dp \right)^{1/p} = \lim_{k \to \infty} \left\| \sum_{i=m+1}^{k} x_i \right\|_p \leq \sum_{i=m+1}^{\infty} \|x_i\|_p$$

for any positive integer $m$. Our claim that follows, because $\sum_{i=0}^{\infty} \|x_i\|_p < \infty$ implies that $\sum_{i=m+1}^{\infty} \|x_i\|_p \downarrow 0$ as $m \uparrow \infty$ (Exercise 1.10 of Appendix 1).

**Remark 5.2.** The Riesz-Fischer Theorem remains valid in the context of any measure space, that is, $\mathcal{L}^p(X, \Sigma, \mu)$ is a Banach space for any $p \in [1, \infty)$ and measure space $(X, \Sigma, \mu)$. The proof of this fact is identical to that we gave above, except that we should now use Exercise 1.17 instead of the Monotone Convergence Theorem 1. $lacksquare$

---

$^{38}$The core of this result goes back to 1907. It was proved by Frigyes Riesz and Ernst Fischer independently in the case of square-integrable functions.
In the following set of exercises \((X, \Sigma, p)\) stands for an arbitrarily fixed probability space and \(p \in [1, \infty)\).

**Exercise 5.6.** Consider the semimetric \(d_p\) on \(\Sigma\) introduced in Example B.2.10. Use the Riesz-Fischer Theorem to prove that \((\Sigma, d_p)\) is complete.

**Exercise 5.7.** Take any sequence \((x_m)\) in \(L^p(X, \Sigma, p)\) such that \(\|x_m - x\|_p \to 0\). Show that there is a strictly increasing sequence \((m_k)\) of natural numbers such that \(x_{m_k} \to a.s. x\).

**Exercise 5.8.** Take any sequence \((x_m)\) in \(L^p(X, \Sigma, p)\) such that \(x_m \to a.s. x\) for some \(x \in L^p(X, \Sigma, p)\). Show that \(\|x_m - x\|_p \to 0\) iff \(\|x_m\|_p \to \|x\|_p\).

**Exercise 5.9.** \((\text{Orlicz Spaces})\) Let \(\psi\) be a convex and strictly increasing self-map on \(\mathbb{R}_+\) with \(\psi(0) = 0\) and \(\psi(\infty) = \infty\). (Note that \(\psi\) must be continuous.) Let \(L_\psi(X, \Sigma, p)\) stand for the linear space of all random variables on \((X, \Sigma, p)\) such that \(\mathbb{E}(\psi(|x|/\lambda)) < \infty\) for some real number \(\lambda > 0\). We define the real map \(\|\cdot\|_\psi\) on \(L_\psi(X, \Sigma, p)\) by

\[
\|x\|_\psi := \inf \left\{ \lambda > 0 : \mathbb{E} \left( \psi \left( \frac{|x|}{\lambda} \right) \right) \leq 1 \right\}.
\]

(a) Show that \(\|\cdot\|_\psi\) is a norm on \(L_\psi(X, \Sigma, p)\). (Here \(\|\cdot\|_\psi\) is said to be a **Luxemburg norm**.) When endowed with this norm, \(L_\psi(X, \Sigma, p)\) is said to be a **Orlicz space**.

(b) Show that \(L_\psi(X, \Sigma, p)\) is a normed linear subspace of \(L^1(X, \Sigma, p)\).

(c) For any \(p \geq 1\), show that \(L_\psi(X, \Sigma, p)\) reduces to \(L^p(X, \Sigma, p)\), provided that \(\psi(t) = t^p\) for all \(t \geq 0\).

(d) The **heart** of \(L_\psi(X, \Sigma, p)\) is the set \(H_\psi(X, \Sigma, p)\) of all \(x \in L_\psi(X, \Sigma, p)\) with \(\mathbb{E}(\psi(|x|/\lambda)) < \infty\) for every \(\lambda > 0\). Show that \(H_\psi(X, \Sigma, p)\) is a closed linear subspace of \(L_\psi(X, \Sigma, p)\).

(e) Prove that \(L_\psi(X, \Sigma, p)\) is a Banach space.

**Exercise 5.10.** (This exercise assumes familiarity with the material covered in Section 2.5.) \(BL^1,Y(X, \Sigma, p)\) is complete for any Banach space \(Y\) and probability space \((X, \Sigma, p)\). Prove!

\(\mathcal{L}^\infty\) **Spaces**

Consider the probability space \(([0,1], B[0,1], \ell)\) and define the random variable \(x\) on this space by letting \(x(\omega) := 1/\omega\) if \(\omega \in \mathbb{Q}\) and \(x(\omega) := 1\) otherwise. This random variable is unbounded. Yet, \(x\) is bounded almost surely, or put differently, it is unbounded only “where it doesn’t matter anyway,” that is, on a set of probability zero. (For instance, \(\mathbb{E}(x) = 1\), right?) Especially for probabilistic applications, this weaker notion of boundedness is every bit as useful as the standard notion of boundedness.

Let us formalize the matter in the context of an arbitrary measure space \((X, \Sigma, \mu)\). We say that a map \(f\) in \(L^0(X, \Sigma)\) is **essentially bounded** (with respect to \(\mu\)) if there exists a number \(K > 0\) and a set \(S \in \Sigma\) with \(\mu(S) = 0\) such that \(|x(\omega)| \leq K\) for all \(\omega \in X \setminus S\). The set of all such random variables is denoted by \(\mathcal{L}^\infty(X, \Sigma, \mu)\). Clearly, \(B(X) \cap L^0(X, \Sigma) \subseteq \mathcal{L}^\infty(X, \Sigma, \mu)\), but the latter cannot be normed by the sup-norm (as the previous example shows). Instead we use the following modification of the sup-norm for \(\mathcal{L}^\infty(X, \Sigma, \mu)\):

\[
\|x\|_\infty := \inf\{K > 0 : \mu\{|x| \leq K\} = 1\}.
\]

(Notice that \(\|x\|_\infty\) reduces to the usual sup-norm on \(B(X) \cap L^0(X, \Sigma)\), so this notation is natural.) Identifying, again, those random variables that are equal to each other almost surely, and using \(\|\cdot\|_\infty\), we can view \(\mathcal{L}^\infty(X, \Sigma, \mu)\) as a normed linear space. This space too is Banach. (The proof of this is actually easier than that of the Riesz-Fischer Theorem.) Notice also that a sequence \((f_m)\) in \(\mathcal{L}^\infty(X, \Sigma, \mu)\) converges to a map \(f\) in \(\mathcal{L}^\infty(X, \Sigma, \mu)\) iff \((f_m)\) converges to \(f\) almost uniformly. Another way of saying this is that we have \(\|f_m - f\|_\infty \to 0\) iff there exist a sequence \((g_m)\) and random variable \(g\) in \(L^0(X, \Sigma)\) such that \(f =_{a.s.} g\), \(f_m =_{a.s.} g_m\) for each \(m\), and \(g_m \to g\) uniformly.
Exercise 5.11.\textsuperscript{H} Show that $L^\infty(X, \Sigma, \mu)$ is a Banach space.

Exercise 5.12. Take any $f$ in $L^\infty(X, \Sigma, \mu)$ which is $q$-integrable for some $q \in [1, \infty)$. Use Exercise 1.17 to prove that $\|f\|_p \to \|f\|_\infty$.

**Approximation in $L^p$ Spaces**

Clearly, every simple random variable on a probability space is $p$-integrable, $p \geq 1$. In fact, every $p$-integrable random variable on a probability space can be approximated in the $L^p$-norm by simple random variables to any desired degree of accuracy.

**Proposition 5.1.** Let $(X, \Sigma, \mu)$ be a probability space. For any real number $p \geq 1$, the set of all simple random variables on $(X, \Sigma)$ is dense in $L^p(X, \Sigma, \mu)$.

**Proof.** Take any $p \geq 1$ and $x \in L^p(X, \Sigma, \mu)$. By Proposition C.1.10, we can find two sequences of nonnegative simple random variables, say, $(u_m)$ and $(v_m)$, on $(X, \Sigma)$ such that $u_m \uparrow x^+$ and $v_m \uparrow x^-$. Set $z_m := u_m - v_m$ for each $m$. Then, $(z_m)$ is a sequence of simple random variables on $(X, \Sigma)$ such that $z_m \to x$ and $|z_m| \leq |x|$ (because $|u_m - v_m| \leq u_m + v_m \leq x^+ + x^-$). Moreover, $|x - z_m|^p$ is bounded above by an integrable random variable on $(X, \Sigma, \mu)$, because

$$|x - z_m|^p \leq (|x| + |z_m|)^p \leq 2^p |x|^p$$

and $\int_X |x|^p \, d\mu < \infty$. So, by the Dominated Convergence Theorem, $\|x - z_m\|_p \to 0$. \hfill \qed

In the case of an arbitrary measure space, we need to modify this proposition suitably.

Exercise 5.13. Let $(X, \Sigma, \mu)$ be a measure space and $p \in [1, \infty)$. Let $S$ stand for the set of all random variables $f$ on $(X, \Sigma)$ such that $\mu\{f \neq 0\} < \infty$. Show that $S$ is dense in $L^p(X, \Sigma, \mu)$.

Exercise 5.14. Let $(X, \Sigma, \mu)$ be a $\sigma$-finite measure space such that $\Sigma$ is countably generated. Show that $L^p(X, \Sigma, \mu)$ is separable for any $p \in [1, \infty)$. (In particular, $L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$ is separable for any $p \in [1, \infty)$.)

Exercise 5.15.\textsuperscript{H} Show that $L^\infty([0, 1], \mathcal{B}[0, 1], \ell)$ is not separable.

A continuous function (on a Borel probability space) need not be integrable, but a continuous and bounded function is surely $p$-integrable for any $p \geq 1$. Indeed, for any Borel probability space $(X, \mathcal{B}(X), \mathbf{p})$, we have $C_b(X) \subseteq L^p(X, \mathcal{B}(X), \mathbf{p})$ because $C(X) \subseteq L^0(X, \mathcal{B}(X))$ and

$$\int_X |x|^p \, d\mathbf{p} \leq \|x\|_\infty^p < \infty \quad \text{for every } x \in C_b(X) \text{ and } p \geq 1.$$

We will next show that $C_b(X)$ is in fact a dense subset of $L^p(X, \mathcal{B}(X), \mathbf{p})$.

**Lemma 5.2.** Let $X$ be a metric space and $S$ a nonempty closed subset of $X$. Then, there exists a sequence $(\varphi_m)$ of nonnegative, continuous and bounded real maps on $X$ such that $\varphi_m \to 1_S$.

**Proof.** For any positive integer $m$, let $A_m := X \setminus N_{1/m}(S)$, and note that $A_m \in \mathcal{C}_X$. Thus, by Urysohn’s Lemma (Appendix 2), there is a continuous $\varphi_m : X \to [0, 1]$ with $\varphi_m|S = 1$ and $\varphi_m|A_m = 0$. It is readily checked that $\varphi_m \to 1_S$. \hfill \square

In passing, we note that Lemma 5.2 shows that the set of all continuous and bounded functions on a metric space is in general not complete under the $L^p$-norm. (This is so whenever there is at least one discontinuous simple random variable in the associated space.)
**Lemma 5.3.** Let $X$ be a metric space, $x$ a simple random variable on a probability space $(X, \mathcal{B}(X), p)$ and $1 \leq p < \infty$. Then, there is a sequence $(\varphi_m)$ in $C_b(X)$ such that $\|\varphi_m - x\|_p \to 0$.

**Proof.** As linear combinations of continuous functions remain continuous, we may assume that $x$ is of the form $\mathbf{1}_A$ for some Borel subset $A$ of $X$. Fix any $m \in \mathbb{N}$. By regularity of $p$ (Exercise B.2.14), there is an $S \in \mathcal{C}_X$ with $S \subseteq A$ and $p(A) - p(S) < 1/m^p$. Then, $\|\mathbf{1}_A - \mathbf{1}_S\|_p < 1/m$, while, by Lemma 5.2, there is a $\varphi \in C_b(X)$ with $\|\varphi - \mathbf{1}_S\|_p < 1/m$. By Minkowski’s Inequality, then, $\|\varphi - \mathbf{1}_A\|_p < 2/m$. It follows that $\|\varphi_m - x\|_p \to 0$, as we sought.

**The Approximation Theorem for $L^p$.** Let $X$ be a metric space, $p \in \Delta(X)$ and $1 \leq p < \infty$. Then, $C_b(X)$ is dense in $L^p(X, \mathcal{B}(X), p)$.

**Proof.** Apply Lemma 5.3 and Proposition 5.1.

**Some Comments on $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$**

The approximation theorem we proved above is not valid in the case of infinite measure spaces. For instance, $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$ does not even contain $C_c(\mathbb{R})$; after all, no nonzero constant function belongs to $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$. There is nevertheless a useful approximation theorem for $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$ which we will need later. First, let us agree to say that a real self-map $f$ on $\mathbb{R}$ is a **continuous map with compact support** if it is continuous and the closure of the set $\{t \in \mathbb{R} : f(t) \neq 0\}$ is compact in $\mathbb{R}$. We denote the collection of all such functions by $C_c(\mathbb{R})$. An easy application of Weierstrass’ Theorem shows that $C_c(\mathbb{R}) \subseteq L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$. We now show that $C_c(\mathbb{R})$ is dense in $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$, that is, any integrable self-map on $\mathbb{R}$ can be approximated by a member of this collection (with respect to $\| \cdot \|_1$) to any desired degree of accuracy.

**The Approximation Theorem for $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$.** $C_c(\mathbb{R})$ is dense in $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$.

To streamline the basic argument we will introduce two auxiliary concepts. First, we say that a function of the form $\sum_{i \in [m]} \mathbf{1}_{A_i}$ is a **simple Borel function** if $m$ is a positive integer, and we have $(a_i, A_i) \in \mathbb{R} \times \mathcal{B}(\mathbb{R})$ with $\ell(A_i) < \infty$ for each $i \in [m]$. In turn, we say that this function is very simple if each $A_i$ here is a (bounded) interval. Let us denote the set of all simple Borel functions by $\mathcal{S}_0$ and that of all very simple Borel functions by $\mathcal{S}_1$. Obviously, $\mathcal{S}_1 \subseteq \mathcal{S}_0 \subseteq L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$.

Our first observation is that $\mathcal{S}_0$ is dense in $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$. The proof of this is identical to that of Proposition 5.1 except we now use the Lebesgue Convergence Theorem (Exercise 1.19) instead of the Dominated Convergence Theorem. Next, we claim that $\mathcal{S}_1$ is dense in $\mathcal{S}_0$. A moment’s reflection (which appeals to the subadditivity of $\| \cdot \|_1$) shows that this claim will be proved if we can show that any indicator function $\mathbf{1}_A$, where $A$ is a Borel subset of $\mathbb{R}$ with $\ell(A) < \infty$, can be approximated by very simple Borel functions to any desired degree of accuracy. But, by the Approximation Lemma for $\ell$ (Exercise B.3.21) and Lemma B.1.1, for any such $A$ and $\varepsilon > 0$, there exist finitely many (disjoint) right-semiclosed intervals, say, $I_1, \ldots, I_m$, such that $\ell(A \Delta \bigcup^m I_i) < \varepsilon$, which means $\|\mathbf{1}_A - \sum_{i \in [m]} \mathbf{1}_{I_i}\|_1 < \varepsilon$. We conclude: $\mathcal{S}_1$ is dense in $\mathcal{S}_0$.

We now know that $\mathcal{S}_1$ is dense in $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$. To complete our proof, therefore, it is enough to show that any given very simple Borel function can be approximated by a continuous self-map on $\mathbb{R}$ with compact support. As linear combinations of the elements of $C_c(\mathbb{R})$ remain in $C_c(\mathbb{R})$, it is again enough to prove this using only the indicator functions. To this end, take any bounded interval $I$ and fix an arbitrary $\varepsilon > 0$. Pick any two real numbers $a$ and $b$ such that the closure of $I$ is contained in $(a, b)$ and $b - a - \ell(I) < \varepsilon$. Clearly, we can construct (by interpolation) a piecewise continuous self-map $f$ on $\mathbb{R}$ such that $\|\mathbf{1}_A - f\|_1 < \varepsilon$ and $\{t \in \mathbb{R} : f(t) \neq 0\} = (a, b)$. The proof of the Approximation Theorem for $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \ell)$ is now complete.

We conclude this section by noting that we can also define $p$-integrability for self-maps on $\mathbb{R}$ (with respect to $\ell$) in the obvious way, and indeed, everything we have just put on record for integrable self-maps on $\mathbb{R}$ would carry over to such maps.
5.2 $\mathcal{L}^p$-Boundedness and Uniform Integrability

For any given real number $p \geq 1$, we say that a set $\mathcal{X}$ of random variables on a given probability space is $\mathcal{L}^p$-bounded if either it is empty or

$$\sup\{\mathbb{E}(|x|^p) : x \in \mathcal{X}\} < \infty.$$  

(An easy application of Jensen's Inequality shows that if $\mathcal{X}$ is $\mathcal{L}^p$-bounded, then it is $\mathcal{L}^q$-bounded for every $q$ in $[1, p]$.) The objective of this section is to identify the connection between the notions of $\mathcal{L}^p$-boundedness and uniform integrability. We will see that the nature of this connection depends on whether we have $p = 1$ or $p > 1$.

Let us first note that uniform integrability is more demanding than $\mathcal{L}^1$-boundedness. Indeed, if $\mathcal{X}$ is a uniformly integrable set of random variables on a probability space $(X, \Sigma, \mathbb{P})$, then either $\mathcal{X}$ is empty, or there is a positive number $a$ such that

$$\int_{\{|x| > a\}} |x| \, d\mathbb{P} < 1$$

for each $x \in \mathcal{X}$, which implies $\mathbb{E}(|x|) \leq 1 + a$ for each $x \in \mathcal{X}$. Thus:

**Proposition 5.4.** Let $\mathcal{X}$ be a nonempty set of random variables on a probability space $(X, \Sigma, \mathbb{P})$. If $\mathcal{X}$ is uniformly integrable, then it is $\mathcal{L}^1$-bounded.

The converse of Proposition 5.4 is false. To give a counterexample, consider the map $x_m := m 1_{[0, 1/m)}$ where $m \in \mathbb{N}$ and $1_{[0, 1/m)}$ is the indicator function of $[0, 1/m)$ on $[0, 1)$. Then, $\{x_1, x_2, \ldots\}$ is an $\mathcal{L}^1$-bounded set of random variables on the probability space $([0, 1), \mathcal{B}[0, 1), \ell)$, but this collection is not uniformly integrable, because for any real number $a > 0$ and integer $m > a$, we have $\mathbb{E}(|x_m| 1_{\{|x_m| > a\}}) = 1$.

Curiously, however, any $\mathcal{L}^p$-bounded set of random variables is uniformly integrable, provided that $p > 1$. This readily follows from Exercise 2.33.(a), but it is also easy to prove directly.

**Proposition 5.5.** Let $\mathcal{X}$ be a set of random variables on a probability space $(X, \Sigma, \mathbb{P})$. If $\mathcal{X}$ is $\mathcal{L}^p$-bounded for some $p > 1$, then it is uniformly integrable.

**Proof.** Take any real number $p > 1$, fix an $\varepsilon > 0$, and assume that $\mathcal{X}$ is nonempty and $\mathcal{L}^p$-bounded. Notice that, for any $(\omega, x) \in X \times \mathcal{X}$ and a real number $a > 0$ with $x(\omega) > a$, we have $x(\omega)^{1-p} < a^{1-p}$. Consequently, for any $a > 0$ and $x \in \mathcal{X}$,

$$\int_{\{|x| > a\}} |x| \, d\mathbb{P} \leq \int_{\{|x| > a\}} |x|^p a^{1-p} \, d\mathbb{P} \leq a^{1-p} b,$$

and hence,

$$\sup \left\{ \int_{\{|x| > a\}} |x| \, d\mathbb{P} : x \in \mathcal{X} \right\} \leq a^{1-p} b,$$
where \( b := \sup \{ \mathbb{E}(|x|^p) : x \in \mathcal{X} \} \), which is finite by hypothesis. But since \( p > 1 \), the map \( a \mapsto a^{1-p}b \) is strictly decreasing and continuous on \( \mathbb{R}_{++} \) with \( \sup \{ a^{1-p}b : a > 0 \} = 0 \), so there is an \( a > 0 \) large enough so that \( a^{1-p}b \leq \varepsilon \), and we are done. 

**Insight.**

\[
\text{\( L^p \)-boundedness} \quad p > 1 \quad \Rightarrow \quad \text{uniform integrability} \quad \Rightarrow \quad \text{\( L^1 \)-boundedness}
\]

**Exercise 5.16.** Let \( (x_m) \) be a sequence of random variables on a probability space \( (X, \Sigma, \mathbb{P}) \) such that both \( \sup |\mathbb{E}(x_m)| \) and \( \sup \mathcal{V}(x_m) \) are finite. Prove that \( (x_m) \) is uniformly integrable.

While \( L^1 \)-boundedness of a random sequence does not guarantee its uniform integrability, it is worth noting that \( L^1 \)-convergence does. That is, every convergent sequence in \( L^1(X, \Sigma, \mathbb{P}) \) is uniformly integrable, as we prove next.

**Proposition 5.6.** Let \( x, x_1, x_2, \ldots \) be integrable random variables on a probability space \( (X, \Sigma, \mathbb{P}) \) such that \( \mathbb{E}(|x_m - x|) \to 0 \). Then, \( (x_m) \) is uniformly integrable.

**Proof.** In view of Example 2.12, it is clear that \( (x_m) \) is uniformly integrable iff so is \( (x_m + x) \). Consequently, we may assume here that \( x = 0 \) without loss of generality. Take any \( \varepsilon > 0 \). By hypothesis, there is a positive integer \( M \) such that \( \mathbb{E}(|x_m|) < \varepsilon \) for each \( m \geq M \). Then, \( \mathbb{E}(|x_m|1_{\{|x_m|>a\}}) \leq \mathbb{E}(|x_m|) < \varepsilon \) for every \( a > 0 \) and \( m \geq M \).

But, as we noted in Example 2.10, integrability of \( x_1, \ldots, x_m \) imply that there is an \( a > 0 \) large enough that \( \mathbb{E}(|x_m|1_{\{|x_m|>a\}}) < \varepsilon \) for each \( m \in [M] \). Combining these two observations completes our proof. 

**Insight.**

\[
\text{\( L^1 \)-convergence} \quad \Rightarrow \quad \text{uniform integrability} \quad \Rightarrow \quad \text{\( L^1 \)-boundedness}
\]

It is plain that the converse of Proposition 5.6 is false. There are, however, important classes of random sequences for which \( L^1 \)-convergence and uniform integrability are identical properties. We will encounter one such case in Chapter I.

### 5.3 Almost Sure Convergence vs. \( L^p \)-Convergence

One of the primary objectives of probability (limit) theory is understanding the convergence behavior of random sequences. There are various modes of convergence that one might use in this regard, and we have already encountered one of the most important of such modes, namely, almost sure convergence. But our analysis of \( L^p \) spaces furnishes us with another notion of convergence. After all, any \( L^p \) space is a semimetric space (relative to the semimetric induced by \( \| \cdot \|_p \)), so there is a natural notion of convergence for sequences in it, which we call \( L^p \)-convergence. In general, almost sure convergence and \( L^p \)-convergence are logically independent concepts. We
illustrate this in the case of $L^1$-convergence (which, by the way, is sometimes referred to as convergence in the mean), but an easy modification of these examples shows that the same is true for $L^p$-convergence for any $p \geq 1$.

**Example 5.1.** Almost sure convergence does not imply $L^1$-convergence. Consider the sequence $(x_m)$ of random variables on the probability space $([0,1],\mathcal{B}[0,1],\ell)$ where $x_m$ equals $m$ on $[0,\frac{1}{m}]$ and 0 elsewhere on $[0,1]$. Then, $x_m(\omega) \to 0$ for each $\omega \in (0,1]$, and hence, $x_m \to_{a.s.} 0$. But $\|x_m\|_1 = 1$ for each $m$. \qed

**Example 5.2.** $L^1$-convergence does not imply almost sure convergence. Consider

$$(x_m) := (1_{[0,1)}, 1_{[0,1/2)}, 1_{[1/2,1)}, 1_{[0,1/3)}, 1_{[1/3,2/3)}, \ldots),$$

which is a sequence in $L^1([0,1),\mathcal{B}[0,1),\ell)$. Clearly, we have $\mathbb{E}(x_m) \to 0$, that is, $\|x_m\|_1 \to 0$. But $(x_m(\omega))$ does not converge to a real number for any $\omega$ in $[0,1)$. \qed

It turns out that the culprit behind Example 5.1 (but not of Example 5.2) is the sequence $(x_m)$ being not uniformly integrable. The final result of this section shows that almost sure convergence does imply $L^1$-convergence for uniformly integrable sequences. This fact is frequently used to deduce results about “convergence in the mean” from the results that are already known for “almost sure convergence.”

**Proposition 5.7.** Let $(x_m)$ be a uniformly integrable sequence of random variables on a probability space $(X,\Sigma,\mathbb{P})$ such that $x_m \to_{a.s.} x$ for some $x \in L^0(X,\Sigma)$. Then, $x$ is integrable and $\mathbb{E}(|x_m - x|) \to 0$.

**Proof.** By Fatou’s Lemma, we have

$$\int_X |x| \, d\mathbb{P} \leq \liminf \left( \int_X |x_m| \, d\mathbb{P} \right) \leq \sup \{ \mathbb{E}(|x_1|), \mathbb{E}(|x_2|), \ldots \},$$

so the integrability of $x$ follows from the fact that $(x_m)$ is $L^1$-bounded (Proposition 5.4). Next, for any real number $a > 0$, consider the continuous self-map $g_a$ on $\mathbb{R}$ with $g_a(t) := -a$ if $t < -a$, $g_a(t) := t$ if $t \in [-a, a]$, and $g_a(t) := a$ if $t > a$. Clearly,

$$\mathbb{E}(|x_m - x|) \leq \mathbb{E}(|x_m - g_a(x_m)|) + \mathbb{E}(|g_a(x_m) - g_a(x)|) + \mathbb{E}(|g_a(x) - x|) \quad (26)$$

for each $m$. We wish to examine the sizes of the three terms on the right-hand side of this inequality for large $m$ and $a$. Let $\varepsilon > 0$. By uniform integrability, there is a number $a_1 > 0$ such that

$$\mathbb{E}(|x_m - g_a(x_m)|) \leq \int_{\{|x_m| > a\}} |x_m| \, d\mathbb{P} < \frac{\varepsilon}{3}$$

72
for every positive integer $m$ and $a \geq a_1$. On the other hand, since $x$ is integrable, Lemma 2.7 says that there is a real number $a_2 \geq a_1$ such that

$$\mathbb{E}(|g_a(x) - x|) \leq \int_{\{|x|>a\}} |x| \, dp < \frac{\varepsilon}{3}$$

for every $a \geq a_2$. Finally, notice that $|g_a(x_m) - g_a(x)| \leq |x_m - x|$ for each $m$. Since $x_m \xrightarrow{a.s.} x$, therefore, we have $|g_a(x_m) - g_a(x)| \xrightarrow{a.s.} 0$ for every $a$. But $|g_a(x_m) - g_a(x)| \leq 2a$ for every $a$ and $m$, so by the Dominated Convergence Theorem, $\mathbb{E}(|g_a(x_m) - g_a(x)|) \rightarrow 0$ for every $a$. Then, for any fixed $a \geq a_2$, there exists a positive integer $M$ such that

$$\mathbb{E}(|g_a(x_m) - g_a(x)|) < \frac{\varepsilon}{3} \quad \text{for every } m \geq M.$$

Combining these observations with (26) yields $\mathbb{E}(|x_m - x|) < \varepsilon$ for every $m \geq M$. Since $\varepsilon > 0$ is arbitrary here, we are done. \hfill $\blacksquare$

**Corollary 5.8.** Let $x, x_1, x_2, \ldots$ be integrable random variables on a probability space $(X, \Sigma, p)$ such that $x_m \xrightarrow{a.s.} x$. Then, $\mathbb{E}(|x_m - x|) \rightarrow 0$ if, and only if, $(x_m)$ is uniformly integrable.

**Proof.** Apply Propositions 5.6 and 5.7. \hfill $\blacksquare$

**Exercise 5.17.** (A Generalization of Lemma 2.6) Let $(x_m)$ be a sequence of random variables on a probability space $(X, \Sigma, p)$ such that $(|x_m|)$ is uniformly integrable. Prove that

$$\mathbb{E}(\lim inf x_m) \leq \lim inf \mathbb{E}(x_m) \leq \lim sup \mathbb{E}(x_m) \leq \mathbb{E}(\lim sup x_m).$$

(In view of Example 2.11, this observation generalizes Lemma 2.6.)

**Exercise 5.18.** (A Dominated Convergence Theorem) Let $(X, \Sigma, p)$ be a probability space and $(x_m)$ a sequence in $L^1(X, \Sigma, p)$ such that $x_m \xrightarrow{a.s.} x$ for some random variable $x$ on $(X, \Sigma, p)$. Prove: $(x_m)$ is uniformly integrable iff $x$ is integrable and $\mathbb{E}(x_m) \rightarrow \mathbb{E}(x)$.

## 6 Probability Measures Induced by Expectations

Let $(X, \Sigma)$ be a measurable space. Given any probability measure on $\Sigma$, may view the Lebesgue integral with respect to that measure as a linear functional on a suitable linear subspace of $L^0(X, \Sigma)$. This is one of the main messages of Section 1, which arrives at the “expectation functional” from the “probability measure.” Interestingly, this causality can be reversed in a variety of cases. That is, it is possible to take a certain type of a linear functional $L$ on a suitable linear subspace of $L^0(X, \Sigma)$, and then derive a probability measure on $\Sigma$ such that $L$ is indeed the expectation functional relative to that measure.

We shall formalize this approach in this section in the context of random variables defined on a compact metric space. Not only will this complete our introduction to the Lebesgue integration theory, but it will also introduce one of the most commonly used tools of individual decision theory.
6.1 The Riesz-Radon Representation Theorem

An Axiomatization of the Expectation Functional

Let $X$ be a metric space and $p \in \Delta(X)$. We know from Example C.1.3 that $C(X) \subseteq L^0(X, \mathcal{B}(X))$. We also have $C_b(X) \subseteq L^1(X, \mathcal{B}(X), p)$ because $\|x\|_1 \leq \|x\|_{\infty} < \infty$ for every $x \in C_b(X)$. As compactness of $X$ ensures $C(X) = C_b(X)$, therefore, we may conclude: For any compact metric space $X$,

$$C(X) \subseteq L^1(X, \mathcal{B}(X), p) \quad \text{for every } p \in \Delta(X).$$

(27)

Thus, the expectation functional $x \mapsto \int_X x dp$ is a (well-defined) real map on $C(X)$, provided that $X$ is compact.

This map possesses a number of nice properties. For one thing, it is a linear functional. That is, we have

$$\int_X (ax + y) dp = a \int_X x dp + \int_X y dp \quad \text{for every } a \in \mathbb{R} \text{ and } x, y \in C(X).$$

(28)

It is in fact a positive linear functional, that is,

$$\int_X x dp \geq 0 \quad \text{for every } x \in C(X) \text{ with } x \geq 0.$$  

(29)

Moreover, the expectation functional on $C(X)$ maps $1_X$ to 1, because

$$\int_X 1_X dp = 1.$$  

(30)

This is all good, but is it enough to make the expectation functional a terribly interesting map on $C(X)$? After all, there are probably a zillion other real maps on $C(X)$ that satisfy the properties above. Surely, these alone cannot provide sufficient reasons to view the expectation functional as a special object. Well, as a matter of fact, they do. There do not exist a zillion other such objects. The map $x \mapsto \int_X x dp$, with $p$ being arbitrarily chosen in $\Delta(X)$, is the only one!

Representation of Positive Linear Functionals on $C(X)$

The Riesz-Radon Representation Theorem.\(^{39}\) Let $X$ be a compact metric space and $L$ a positive linear functional on $C(X)$ with $L(1_X) = 1$. Then, there exists a unique Borel probability measure $p$ on $X$ such that

$$L(\varphi) = \int_X \varphi dp \quad \text{for every } \varphi \in C(X).$$

(31)

\(^{39}\)Time for some name dropping. Frigyes Riesz has proved this result in 1909 in the special case where $X = [0,1]$, and using the notion of Stieltjes integral (Chapter E), not the Lebesgue integral. In 1913, Johann Radon extended the result to the case where $X$ is an arbitrary compact subset of a Euclidean space – he was the first mathematician who used the Lebesgue integral in the representation. Various generalizations have been obtained since then. In particular, the formulation I present here was given first by Stefan Banach in 1937. (For a comprehensive account of the history of this result and its importance thereof, see Gray (1984).)
This is a thing of beauty, to be sure. But how does one prove something like this? Let’s see. Let $L$ be a positive linear functional on $C(X)$ with $L(1_X) = 1$, and suppose that the theorem is actually true, that is, “assume” that we can indeed find a unique Borel probability measure $p$ on $X$ which satisfies (31). Then, for each Borel subset $S$ of $X$, we may be able to get a sense of $p(S)$ by choosing a map $\varphi$ in a clever way. For instance, if $\varphi$ is the indicator function $1_S$ on $X$, then we would have

$$L(1_S) = \int_X 1_S d\mu = p(S).$$

This looks great, why don’t we then just define $p(S) := L(1_S)$ for all $S \in \mathcal{B}(X)$? Well, as promising as it is, this approach would not quite do, because $L(1_S)$ may well be – and it often is – undefined. After all, the map $1_S$ need not be continuous, so $L$, whose domain is $C(X)$, is not primed to assign a value to this function. But it would really be a pity to let go of this choice of $p$: How about we approximate $1_S$ with continuous functions and designate as the probability of $S$ the limit of the $L$ values of the approximating functions? This suggests defining something like

$$p(S) := \inf \{ L(\varphi) : \varphi \in C(X) \text{ and } \varphi \geq 1_S \} \quad \text{for every } S \in \mathcal{B}(X).$$

This is better, but it still won’t do as it is, for this definition does not even guarantee that $p$ is finitely additive. Yet, and this is crucial, this definition does give us additivity across closed subsets of $X$ (but this requires proof). In turn, this property can be used to find the “probability” of open sets. It will then be easier to get our hands on the Borel probability measure that we seek.

We will flesh out this idea, which, in great many aspects parallels the one we used to prove Carathéodory’s Extension Theorem, in just a little bit.

In the following set of exercises, $X$ stands for an arbitrarily fixed compact metric space.

**Exercise 6.1. (Uniqueness in the Riesz-Radon Theorem)** If $p$ and $q$ are two Borel probability measures on $X$ such that $\int_X \varphi d\mu = \int_X \varphi d\nu$ for every $\varphi \in C(X)$, then $p = q$.

(a) Use Urysohn’s Lemma to show that, for any $S \in \mathcal{C}_X$, there is a sequence $(\varphi_m)$ in $C(X)$ with $\varphi_m \to 1_S$.

(b) Use part (a) and Proposition B.4.1 to prove the above uniqueness claim.

**Exercise 6.2.** Let $X$ denote the class of all increasing real functions on $[0, 1]$ and let $Y$ stand for the set of all positive linear functionals on $C[0, 1]$. Show that there is a homeomorphism between $X$ and $Y$ which is additive and positively homogeneous.

**Exercise 6.3.** Take any sequence $(\omega_m)$ in $[0, 1]$ and let $(\lambda_m) \in \mathbb{R}_+^\infty$ be such that $\lambda := \sum_{i=1}^{\infty} \lambda_i \in \mathbb{R}_+$. Show that there is a $\mu \in \Delta[0, 1]$ such that

$$\int_{[0,1]} x d\mu = \frac{1}{\lambda} \sum_{i=1}^{\infty} \lambda_i x(\omega_i) \quad \text{for every } x \in C[0, 1].$$

**Exercise 6.4. (Jordan Decomposition Theorem)** Take any continuous linear functional $L$ on $C(X)$. The objective of this exercise is to prove that there exist two positive linear functionals
\[ L_+ \text{ and } L_- \text{ on } C(X) \text{ such that } L = L_+ - L_-.
\]

(a) Let \( C_+(X) := \{ \varphi \in C(X) : \varphi \geq 0 \} \) and define \( K : C_+(X) \to \mathbb{R} \) by
\[
K(\varphi) := \sup \{ L(\phi) : \varphi \geq \phi \in C_+(X) \}.
\]

Show that \( K \) is well-defined, \( K \geq 0 \) and \( K \geq L|_{C_+(X)} \).

(b) Show that \( K(a\varphi + \phi) = aK(\varphi) + K(\phi) \) for every \( a \geq 0 \) and \( \varphi, \phi \in C_+(X) \).

(c) Define \( L_+ : C(X) \to \mathbb{R} \) by
\[
L_+(\varphi) := K(\varphi + \|\varphi\|_{\infty} \, 1_X) - K(\|\varphi\|_{\infty} \, 1_X),
\]
and let \( L_- := L_+ - L \). Show that both \( L_+ \) and \( L_- \) are positive linear functionals on \( C(X) \).

It is plain that we can combine the Jordan Decomposition Theorem with the Riesz-Radon Representation Theorem in the following way:

**Corollary.** For any compact metric space \( X \) and any continuous linear functional \( L \) on \( C(X) \), there exist two Borel probability measures \( p \) and \( q \) on \( X \) such that
\[
L(\varphi) = \int_X \varphi \, dp - \int_X \varphi \, dq \quad \text{for all } \varphi \in C(X).
\]

**Exercise 6.5.** Let \( L \) be a continuous linear functional on \( C(X) \).

(a) Prove that there exist a real number \( \theta \geq L(1_X) \) and two Borel probability measures \( p \) and \( q \) on \( X \) such that
\[
L(\varphi) = \theta \int_X \varphi \, dp - (\theta - L(1_X)) \int_X \varphi \, dq \quad \text{for all } \varphi \in C(X).
\]

(b) Recall how \( C(X)^* \) is normed, and prove that \( ||L||^* = 2\theta - L(1_X) \).

**Proof of the Riesz-Radon Representation Theorem**

We define the map \( \rho : C_X \to \mathbb{R} \) by
\[
\rho(C) := \inf \{ L(\varphi) : \varphi \in C(X) \text{ and } \varphi \geq 1_C \}.
\]

Obviously, \( \rho(\emptyset) = 0 \) and \( \rho(X) = 1 \). On the other hand, positivity and linearity of \( L \) together imply that \( L \) is increasing. Since \( L(1_X) = 1 \), therefore, \( \rho(C_X) \subseteq [0, 1] \) and \( \rho \) is \( \geq \)-increasing. Here are two other observations about \( \rho \).

**Claim 1.** \( \rho(C \cup D) \leq \rho(C) + \rho(D) \) for any closed subsets \( C \) and \( D \) of \( X \).

\textbf{Proof.} This is an easy \( 2\varepsilon \) argument; we leave it as an exercise. \( \|
\)

**Claim 2.** \( \rho(C \cup D) = \rho(C) + \rho(D) \) for any disjoint closed subsets \( C \) and \( D \) of \( X \).

\textbf{Proof.} Take any \( \varphi \in C(X) \) with \( \varphi \geq 1_{C \cup D} \), and use Urysohn’s Lemma to find a continuous \( \phi : X \to [0, 1] \) with \( \phi|_C = 0 \) and \( \phi|_D = 1 \). Then, \( (1 - \phi) \varphi \geq 1_C \) and \( \phi \varphi \geq 1_D \), so, by additivity of \( L \),
\[
L(\varphi) = L((1 - \phi) \varphi + \phi \varphi) = L((1 - \phi) \varphi) + L(\phi \varphi) \geq \rho(C) + \rho(D).
\]

In view of the arbitrary choice of \( \varphi \in C(X) \), we may thus conclude that \( \rho(C \cup D) \geq \rho(C) + \rho(D) \).

Now apply Claim 1. \( \|
Next, we use this measure-like function $\rho$ to estimate the “probability” of open subsets of $X$ from below. More precisely, we define the map $\tau : \mathcal{O}_X \to [0, 1]$ by

$$\tau(O) := \sup \{\rho(C) : C \in \mathcal{C}_X \text{ and } C \subseteq O\}. $$

The corresponding properties of $\rho$ entail that $\tau(\emptyset) = 0$, $\tau(X) = 1$, and $\tau$ is $\preceq$-increasing. The following is a consequence of Claim 1.

**Claim 3.** Let $O$ and $U$ be open subsets of $X$. Then,

$$\tau(O \cup U) \leq \tau(O) + \tau(U). \tag{32}$$

**Proof.** Take any $C \in \mathcal{C}_X$ with $C \subseteq O \cup U$. Then, $C \cap O$ and $C \cap U$ are disjoint closed subsets of $X$. Clearly, we can find two open sets $O'$ and $U'$ in $X$ that contain $C \cap O$ and $C \cap U$, respectively. (How?) Define $S := C \cap O'$ and $T := C \cap U'$. It is readily verified that $S$ and $T$ are closed subsets of $X$ that are contained in $O$ and $U$, respectively. Moreover, we have $C = S \cup T$. By Claim 1, therefore,

$$\rho(C) \leq \rho(S) + \rho(T) \leq \tau(O) + \tau(U).$$

As $C$ is an arbitrary closed subset of $X$ with $C \subseteq O \cup U$ here, we are done. ||

**Claim 4.** Let $(O_m)$ be a sequence in $\mathcal{O}_X$. Then,

$$\tau \left( \bigcup_{i=1}^{\infty} O_i \right) \leq \sum_{i=1}^{\infty} \tau(O_i).$$

**Proof.** Fix any $C \in \mathcal{C}_X$ with $C \subseteq \bigcup_{i=m}^{\infty} O_i$. Since $X$ is compact, so is $C$. Therefore, there is a positive integer $m$ such that $C \subseteq \bigcup_{i=m}^{\infty} O_i$. Then, by definition of $\tau$, and Claim 3,

$$\rho(C) \leq \tau \left( \bigcup_{i \in [m]} O_i \right) \leq \sum_{i \in [m]} \tau(O_i) \leq \sum_{i=1}^{\infty} \tau(O_i).$$

In view of the arbitrariness of the choice of $C$, and definition of $\tau$, we are done. ||

Let us now consider the map $Q : 2^X \to [0, 1]$ where

$$Q(S) := \inf \{\tau(O) : O \in \mathcal{O}_X \text{ and } S \subseteq O\}. $$

(Note. $Q$ is called the outer measure induced by $\tau$.\textsuperscript{40}) The corresponding properties of $\tau$ entail that $Q(\emptyset) = 0$, $Q(X) = 1$, and $Q$ is $\preceq$-increasing. Moreover, by definition, $Q$ agrees with $\tau$ on open sets, that is, $Q|_{\mathcal{O}_X} = \tau$.

**Claim 5.** For any sequence $(S_m)$ in $2^X$, we have

$$Q \left( \bigcup_{i=1}^{\infty} S_i \right) \leq \sum_{i=1}^{\infty} Q(S_i).$$

\textsuperscript{40}I will define $p$ as the restriction of $Q$ to $\mathcal{B}(X)$ below, so this definition is only sensible. I don’t define $p$ right away this way, because it would then be much harder to establish the $\sigma$-additivity of $p$. Instead, I shall show that $Q$ is $\sigma$-additive on a suitable $\sigma$-algebra that includes $\mathcal{B}(X)$. This is exactly how Carathéodory’s Extension Theorem was proved earlier.
Proof. Take any sequence \((S_m)\) in \(2^X\), and fix an arbitrary real number \(\varepsilon > 0\). By definition of \(Q\), for every positive integer \(i\), there is an \(O_i \in \mathcal{O}_X\) with \(S_i \subseteq O_i\) and \(Q(S_i) > \tau(O_i) - 2^{-i} \varepsilon\). But then, by Claim 4,

\[
\sum_{i=1}^{\infty} Q(S_i) > \sum_{i=1}^{\infty} \tau(O_i) + \varepsilon \geq \tau \left( \bigcup_{i=1}^{\infty} O_i \right) + \varepsilon \geq Q \left( \bigcup_{i=1}^{\infty} S_i \right) + \varepsilon.
\]

(The final equality is true, because \(\bigcup^\infty O_i\) is an open set that contains \(\bigcup^\infty S_i\).) As \(\varepsilon > 0\) is arbitrary here, we are done. ||

We now proceed exactly as we did when proving Carathéodory’s Extension Theorem. Define

\[
\Sigma := \{ S \in 2^X : Q(S \cap T) + Q((X\setminus S) \cap T) \leq Q(T) \text{ for all } T \subseteq X \},
\]

and note that, by Claim 5,

\[
\Sigma = \{ S \in 2^X : Q(S \cap T) + Q((X\setminus S) \cap T) = Q(T) \text{ for all } T \subseteq X \}.
\]

Claims 3-5 of Section B.3.6 apply without modification, so we may conclude that \(\Sigma\) is a \(\sigma\)-algebra and \(Q|_{\Sigma}\) is a \(\sigma\)-additive function. Moreover:

Claim 6. \(\mathcal{O}_X \subseteq \Sigma\).

Proof. Let \(O\) be an open subset of \(X\). We wish to show that

\[
T \subseteq X \quad \text{implies} \quad Q(T) \geq Q(O \cap T) + Q((X\setminus O) \cap T).
\]

Take any subset \(T\) of \(X\), and fix \(\varepsilon > 0\) arbitrarily. Let \(U\) be an open subset of \(X\) that contains \(T\). By definition of \(\tau\), there is a \(C \in \mathcal{C}_X\) such that \(C \subseteq O \cap U\) and \(\rho(C) > \tau(O \cap U) - \frac{\varepsilon}{2}\). Similarly, there is a \(D \in \mathcal{C}_X\) with \(D \subseteq U \setminus C\) and \(\rho(D) > \tau(U \setminus C) - \frac{\varepsilon}{2}\). Since \(C\) and \(D\) are disjoint, we can use Claim 2 and the fact that \(Q\) is \(\geq\)-increasing to get

\[
\rho(C \cup D) = \rho(C) + \rho(D) > \tau(O \cap U) + \tau(U \setminus C) - \varepsilon \geq Q(O \cap U) + Q((X\setminus O) \cap U) - \varepsilon \geq Q(O \cap T) + Q((X\setminus O) \cap T) - \varepsilon.
\]

(The second inequality follows from the definition of \(Q\) and the fact that \((X\setminus O) \cap U \subseteq U \setminus C\).) Since \(C \cup D \subseteq U\), and \(\varepsilon > 0\) is arbitrary here, then, the definition of \(\tau\) implies \(\tau(U) \geq Q(O \cap U) + Q((X\setminus O) \cap U)\). In view of the arbitrary choice of \(U\), and the definition of \(Q\), we are done. ||

Since \(\Sigma\) is a \(\sigma\)-algebra, Claim 6 implies that \(\mathcal{B}(X) \subseteq \Sigma\). Consequently, \(p := Q|_{\mathcal{B}(X)} \in \Delta(X)\). It remains to establish (31). We shall do this first for an arbitrary \(\varphi \in \mathcal{C}(X)\) with \(0 \leq \varphi \leq 1\). For any positive integer \(m\), define \(O_i := \{ \varphi > \frac{i-1}{m} \}\) for each \(i = 0, ..., m + 1\), and notice that

\[
X = O_0 \supseteq O_1 \supseteq \cdots \supseteq O_{m+1} = \emptyset.
\]

Define \(\varphi_i : X \to [0,1]\) by

\[
\varphi_i(\omega) := \begin{cases} 
0, & \text{if } \omega \in X \setminus O_i \\
m\varphi(\omega) - (i - 1), & \text{if } \omega \in O_i \setminus O_{i+1} \\
1, & \text{if } \omega \in O_{i+1}
\end{cases}
\]

78
for each \(i \in [m]\). It is easy to check that each \(\varphi_i\) is continuous and we have \(\frac{1}{m}(\varphi_1 + \cdots + \varphi_m) = \varphi\). Furthermore,

\[
p(O_{i+1}) \leq L(\varphi_i) \leq p(O_{i-1}) \quad \text{and} \quad p(O_{i+1}) \leq \int_X \varphi_i d\mathbf{p} \leq p(O_{i-1}),
\]

(verify!) and hence

\[
p(O_{i+1}) - p(O_{i-1}) \leq L(\varphi_i) - \int_X \varphi_i d\mathbf{p} \leq p(O_{i-1}) - p(O_{i+1})
\]

for each \(i \in [m]\). As \(p(O_0) = 1\) and \(p(O_{m+1}) = 0\), then, summing over \(i \in [m]\) yields

\[
\left| \sum_{i \in [m]} L(\varphi_i) - m \int_X \varphi_i d\mathbf{p} \right| \leq 1
\]

As \(\frac{1}{m}(\varphi_1 + \cdots + \varphi_m) = \varphi\), linearity of \(L\) and the Lebesgue integral implies \(\left| L(\varphi) - \int_X \varphi d\mathbf{p} \right| \leq \frac{1}{m}\). Letting \(m \uparrow \infty\) establishes our claim.

Finally, take any \(\varphi \in C(X)\). Since \(X\) is compact, there is a real number \(\alpha > 0\) such that \(\phi := \varphi + \alpha > 0\). Applying what we found in the previous paragraph to the function \(\phi/\|\phi\|_\infty\) completes our proof.

Exercise 6.7. (A Generalization of the Riesz-Radon Theorem) Let \(X\) be a metric space. Assume that for every \(\omega \in X\) there is an open set \(O\) such that \(\omega \in O\) and \(\overline{O}(O)\) is compact. (Such a metric space is called locally compact.) Let \(C_c(X)\) stand for the set of all continuous \(f : X \to \mathbb{R}\) such that the closure of the set \(\{\omega \in X : f(\omega) \neq 0\}\) is compact in \(X\). Let \(L\) be a positive linear functional on \(C_c(X)\) with \(L(1_X) = 1\). Prove that there exists a unique Borel probability measure \(\mathbf{p}\) on \(X\) such that

\[
L(\varphi) = \int_X \varphi d\mathbf{p} \quad \text{for every } \varphi \in C_c(X).
\]

6.2 Choquet’s Theorem

Integral Representation of a Point\(^{41}\)

Let \(X\) be a normed linear space and \(S\) a nonempty compact and convex subset of \(X\). Suppose first that \(X\) is finite dimensional. Then, a classic result of convex analysis, Minkowski’s Theorem, says that every point in \(S\) can be expressed as a convex combination of finitely many extreme points of \(S\). (See Section 1.3 of Appendix 3.) That is, for any \(\omega\) in \(S\), there exist a positive integer \(n\), nonnegative real numbers \(a_1, \ldots, a_n\) with \(\sum_{i \in [n]} a_i = 1\) and \(\nu_1, \ldots, \nu_n\) in \(\text{ext}(S)\) such that \(\omega = \sum_{i \in [n]} a_i \nu_i\). (In fact, \(n\) can be taken to be at most \(\dim(S) + 1\) here, but this is not important for the present discussion.) This statement implies that \(f(\omega) = \sum_{i \in [n]} a_i f(\nu_i)\) for every affine real map \(f\) on \(S\). Conversely, if \(\omega\) and \(\sum_{i \in [n]} a_i \nu_i\) were distinct, we could find an affine real map \(f\) on \(S\) that takes different values on these two vectors (Corollary 1.11 of Appendix 3). Since every affine map on \(S\) is continuous (because \(\dim(S) < \infty\)) we can therefore restate the conclusion of Minkowski’s Theorem as: For every \(\omega\) in \(S\), there is a simple Borel probability measure \(\mathbf{p}\) on \(X\) such that \(\text{supp}(\mathbf{p}) \subseteq \text{ext}(S)\) and \(f(\omega) = \int_S f d\mathbf{p}\) for every affine \(f \in C(S)\).

\(^{41}\) Section 2.3 of Appendix 3 is a prerequisite for this section.
A remarkable 1956 theorem of Gustave Choquet shows that this conclusion remains valid in the context of any normed linear space (except that \( p \) need not be simple in the general case), and determines exactly which property of \( S \) ensures the uniqueness of the found probability measure. In this section, we shall prove the existence part of this theorem, but state its uniqueness part without proof.\(^42\)

**Choquet’s Theorem.** Let \( S \) be a compact and convex subset of a normed linear space \( X \), and \( \omega \in S \). Then, there is a \( p \in \Delta(S) \) such that \( \text{supp}(p) \subseteq \text{ext}(S) \) and

\[
f(\omega) = \int_S f \, dp \quad \text{for every affine } f \in \mathcal{C}(S).
\]

(33)

Needless to say, the requirement \( \text{supp}(p) \subseteq \text{ext}(S) \) is responsible for the non-triviality of this result. After all, without it, we would be done by setting \( p := \delta_\omega \).

**The Krein-Milman Theorem**

Before we prove Choquet’s Theorem, let us show how this result yields one of the most celebrated results of convex analysis almost immediately. The famous Krein-Milman Theorem, which generalizes Minkowski’s Theorem, says that every compact and convex set \( S \) in a normed linear space \( X \) is the closed convex hull of its extreme points, that is, \( S = \text{cl}(\text{co}(\text{ext}(S))) \). (See Section 2.3 of Appendix 3). Let us put \( T := \text{cl}(\text{ext}(S)) \), which is a compact subset of \( S \). As \( \text{cl}(\text{co}(T)) \) is the smallest closed and convex set that contains the closure of \( \text{ext}(S) \), and \( \text{cl}(\text{co}(\text{ext}(S))) \) is such a set, \( \text{cl}(\text{co}(T)) \subseteq \text{cl}(\text{co}(\text{ext}(S))) \). As the converse containment is obviously true, our assertion is: \( S = \text{cl}(\text{co}(T)) \). To prove this, take any \( \omega \) in \( S \), and use Choquet’s Theorem to find a \( p \in \Delta(S) \) such that \( \text{supp}(p) \subseteq T \) and (33) holds. Then, for any continuous affine real map \( f \) on \( S \), we have

\[
f(\omega) = \int_S f \, dp \leq \sup f(T) \leq \sup f(\text{cl}(\text{co}(T))).
\]

Since \( \text{cl}(\text{co}(T)) \) is closed and convex, we can then invoke the Separating Hyperplane Theorem, or more precisely, Corollary 2.15 of Appendix 3, to conclude that \( \omega \in \text{cl}(\text{co}(T)) \), and we are done.

**Proof of Choquet’s Theorem**\(^43\)

Let \( S \) be a compact and convex subset of a normed linear space \( X \), and put

\[
\mathcal{A}(S) := \{ f \in \mathcal{C}(S) : f \text{ is affine} \}.
\]

Clearly, \( \mathcal{A}(S) \) is a normed linear subspace of \( \mathcal{C}(S) \) that contains all constant functions on \( S \). Moreover, as \( S \) is compact, \( \mathcal{C}(S) \), and hence \( \mathcal{A}(S) \), is separable. (Recall Proposition 6.9 and 4.3 of Appendix 2.) Consequently, there is a sequence \( (f_m) \) in \( \mathcal{A}(S) \) such that \( \| f_m \|_\infty = 1 \) for each \( m \), and \( \{ f_1, f_2, \ldots \} \) is dense in \( \{ f \in \mathcal{A}(S) : \| f \|_\infty = 1 \} \). We define

\[
u := \sum_{i=1}^{\infty} 2^{-i} f_i^2.
\]

By the Weierstrass M-Test (Section 2.1 of Appendix 2), \( \nu \) is well-defined and belongs to \( \mathcal{C}(S) \). As each \( f_i^2 \) is convex, and the (pointwise) limit of convex functions is convex, \( \nu \) is a continuous and

---

\(^{42}\)Choquet’s original theorem applies to more general spaces than normed linear spaces (namely, to the case of metrizable linear spaces), and there are many generalizations of it. A splendid reference for “all things Choquet’s Theorem,” is Phelps (2001).

\(^{43}\)The proof I present here is due to Bonsall (1963).
convex real map on $S$. In fact, $u$ is strictly convex. To see this, take any distinct $\omega_1$ and $\omega_2$ in $S$. By Urysohn’s Lemma, there is a continuous real map $\varphi$ on $S$ with $\|\varphi\|_\infty = 1$ and $\varphi(\omega_1) \neq \varphi(\omega_2)$. Then, by the denseness of $\{f_1, f_2, \ldots\}$, we must have $f_j(\omega_1) \neq f_j(\omega_2)$ for at least one $j \in \mathbb{N}$. But this means that $f_j$ is an affine function on $S$ which is not constant on the line segment between $\omega_1$ and $\omega_2$. This implies that $f_j^2$ is strictly convex on that line segment. (Verify!) Then, as $u$ is the sum of a convex function and $2^{-j}f_j^2$, it follows that $u$ is strictly convex on the line segment between $\omega_1$ and $\omega_2$. As we have chosen $\omega_1$ and $\omega_2$ arbitrarily in $S$, we conclude: $u$ is strictly convex.

Consider the subspace

$$ Y := \text{span}(A(S) \cup \{u\}) $$

of $C(S)$, and note that any given element of $Y$ can be written as $f + au$ for a unique $(f, a) \in A(S) \times \mathbb{R}$ (Proposition 2.2 of Appendix 3). We may thus define $K : Y \to \mathbb{R}$ by

$$ K(f + au) := f(\omega) + au^*(\omega) $$

for any $(f, a) \in A(S) \times \mathbb{R}$. (Here $\omega$ is the element of $S$ that we wish to represent as in Choquet’s Theorem, and $u^* : S \to \mathbb{R}$ is the upper envelope of $u$ defined by $u^*(\nu) := \inf\{f(\nu) : f \in A(S)\}$ and $f \geq u$; see Section 2.3 of Appendix 3). It is plain that $K$ is a linear functional on $Y$.

Next, we define $\Phi : C(S) \to \mathbb{R}$ by

$$ \Phi(\varphi) := \varphi^*(\omega). $$

By Proposition 2.16 of Appendix 3, $\Phi$ is subadditive and positively homogeneous. In addition, we claim that $\Phi|_Y \geq K$. To see this, take any $(f, a) \in A(S) \times \mathbb{R}$, and use Proposition 2.16.(f) of Appendix 3 to get $\Phi(f + au) = (f + au)^*(\omega) = f(\omega) + au^*(\omega)$ when $a \geq 0$. On the other hand, if $a < 0$, then $f + au$ is concave, so by Proposition 2.16.(c) of Appendix 3, $\Phi(f + au) = f(\omega) + au^*(\omega)$. Thus, $\Phi(f + au) \geq K(f + au)$, as we claimed.

We may now apply the Hahn-Banach Theorem to find a linear functional $L$ on $C(S)$ such that $\Phi \geq L$ and $L|_Y = K$. The latter condition implies that $L(u) = u^*(\omega)$ and $L(1_S) = 1$ (because both $u$ and $1_S$ belong to $Y$). Moreover, $L$ is a positive linear functional on $C(S)$, because, if $\varphi \in C(S)$ and $\varphi \geq 0$, then $0 \geq (-\varphi)^*(\omega) = \Phi(-\varphi) \geq L(-\varphi) = -L(\varphi)$. (Here, the first inequality follows from the definition of $(-\varphi)^*$ and the fact that the zero function on $S$ is affine and everywhere larger than $-\varphi$.) Consequently, we may apply the Riesz-Radon Representation Theorem to find a $p \in \Delta(S)$ such that

$$ L(\varphi) = \int_S \varphi d\rho \quad \text{for every } \varphi \in C(S). $$

In particular, $\rho$ satisfies (33), because $L(f) = f(\omega)$ for every $f \in A(S)$.

It remains to prove that $\text{supp}(\rho) \subseteq \text{ext}(S)$. To this end, put $E := \{\omega \in S : u(\omega) = u^*(\omega)\}$. Notice that if $\omega_1$ and $\omega_2$ are two distinct elements of $S$, then, as $u$ is strictly convex and $u^*$ is concave (because the upper envelope of every bounded real map is concave),

$$ u(\frac{1}{2}\omega_1 + \frac{1}{2}\omega_2) < \frac{1}{2}u(\omega_1) + \frac{1}{2}u(\omega_2) \leq \frac{1}{2}u^*(\omega_1) + \frac{1}{2}u^*(\omega_2) \leq u^*(\frac{1}{2}\omega_1 + \frac{1}{2}\omega_2). $$

Conclusion: $E \subseteq \text{ext}(S)$. On the other hand, as $L$ is positive and $u^* \geq u$, we have $L(u^*) \geq L(u)$ while, for any $f \in A(S)$ with $f \geq u$, we have $f \geq u^*$, so $f(\omega) = L(f) \geq L(u^*)$, which means $u^*(\omega) \geq L(u^*)$. As $L(u) = u^*(\omega)$, therefore, we find that $L(u^*) = L(u)$, that is, $\int_S (u^*-u) d\rho = 0$. As both $u^*$ and $u$ are continuous, this is possible iff $u^*$ and $u$ coincide on the support of $\rho$, that is, $\text{supp}(\rho) \subseteq E$. It follows that $\text{supp}(\rho) \subseteq \text{ext}(S)$, and our proof is complete.

Exercise 6.8. In the context of the proof above, show that $E = \text{ext}(S)$.

The Uniqueness Part of Choquet’s Theorem

Recall that a simplex in a normed linear space $X$ is a set $S$ which is a base for $\text{cone}(S)$ and $\text{cone}(S)$ is a lattice relative to $\succ$ (Section 1.4 of Appendix 2). A counterpart of Choquet’s Theorem,
also due to Gustave Choquet, says that, a nonempty closed and convex subset $S$ of $X$ is a simplex in $X$ iff there is a unique $p \in \Delta(S)$ that satisfies $\text{supp}(p) \subseteq \text{ext}(S)$ and (33). However, as we will not need this result in the sequel, and because its proof is a bit more involved than the existence part of Choquet’s Theorem, we will not provide a proof for it here.