Chapter 1

Measure and Integral

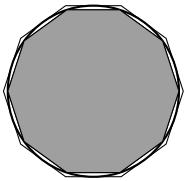
Measure theory stems from the ancient question of measuring areas. It provides definitions of lengths, areas, and volumes under very general circumstances. It is needed for building rigorous foundations of probability theory, and also for modern theory of integration (well, more modern than the 19th century).

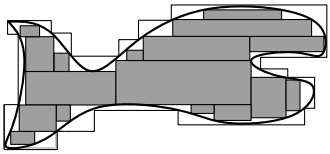
Measure theory is also a basic tool in many discrete and information-theoretic applications. Even if its fine points are seldom encountered in such contexts, it provides solid foundations and indispensable vocabulary.

Defining areas and volumes. What is the area of a possibly complicated geometric figure? This question is definitely not easy.

The area of a rectangle is the product of its sidelengths—that is taken as a basic principle (axiom). It is known that every planar polygon can be dissected into finitely many parts that can be rearranged to make a rectangle, and this can be used to define areas of polygons. However, for the circular disk such a dissection is obviously impossible, and Dehn's famous solution of Hilbert's third problem shows that even some polyhedra, such as the regular tetrahedron, cannot be dissected into finitely many parts that reassemble to a rectangular box.

The area of planar figures such as the disk, and the volume of a tetrahedron and other bodies, can be defined and computed by a limit process, the so-called *method of exhaustion*, which was already known to ancient Greek geometers (Antiphon, Eudoxus). For the disk, for example, one can inscribe and circumscribe regular polygons with larger and larger number of sides, as in the left picture:

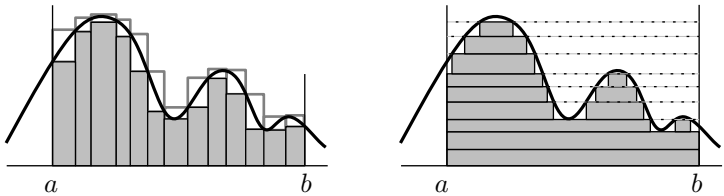




This approach, in more modern form, leads to the notion of *Jordan content* (or Jordan–Peano volume) of a set A, where one approximates A from inside and from outside by a union of finitely many non-overlapping rectangles, as is indicated in the right picture.

The Jordan content of A is defined only if the supremum of the areas of the inner approximations equals the infimum of the outer ones. Unfortunately, this leaves out many natural and important sets of everyday mathematics: for example, the set of all points in the unit square with rational coordinates has undefined Jordan content, because all inner approximations yield 0, while each outer one yields at least 1.

Integration. With the discovery of calculus, many areas and volumes could be computed by integration. Introductory courses almost always present Riemann's definition of integral, where the area under the graph of a nonnegative function f in an interval $[a, b]$ is approximated from the inside and from the outside by finer and finer vertical rectangles, as in the left picture.



Riemann Lebesgue

This is similar to the Jordan content.

Let us note that integrals that can be computed exactly are mostly evaluated symbolically, using antiderivatives, but Riemann's definition is still needed as a formal underpinning of this method.

Here we will introduce a more sophisticated definition of areas and volumes, the *Lebesgue measure* and its abstraction to general *measure spaces*, and a better definition of integral, also due to Lebesgue.

Compared to the Riemann integral, the main conceptual change in Lebesgue's definition is that, instead of cutting vertically, he cuts *horizontally*, or in more abstract terms, instead of subdividing the domain of the integrated function, he subdivides the *range*. A nice metaphor, due to Lebesgue himself, is about counting coins in one's pocket: in the Riemann way, one takes the coins one by one and adds their values; the Lebesgue way is to sort the coins by their denominations and count the number for each denomination.

In the case of integrating a nonnegative function f as in the picture, Lebesgue approximated $\int_a^b f(x)\,\mathrm{d}x$ by the sum of the areas of finitely many horizontal slices. The contribution of the slice between the horizontal lines $y = y_{i-1}$ and $y = y_i$ should be $y_i - y_{i-1}$ times the length of the upper boundary of the slice, i.e., of the set $\{x \in [a,b] : f(x) \geq y_i\}$.

This 1-dimensional set need not be a single interval, or even a finite union of intervals—for a wild function f it can be quite complicated. This is why Lebesgue's definition of integral also needs a satisfactory definition of "length" for very complicated sets in $\mathbb{R}$. (For the Riemann integral this issue does not arise—there we deal only with areas of rectangles.) Similarly, to integrate very general functions on $\mathbb{R}^2$, we need to measure areas of very general planar sets, and so on for higher dimensions.

Fortunately, a suitable notion was already available at the time of Lebesgue's work—although nowadays we call it the Lebesgue measure, it was mostly worked out by Borel and his predecessors. While for the Jordan content we approximate by finite unions of rectangles, for the Lebesgue measure we take *countably many* rectangles (we also approximate only from outside, but that difference is less significant). This seemingly minor change allows one to use the definition for much more general sets, and the resulting notion is better behaved than the Jordan content.

Riemann vs. Lebesgue. For integrating a bounded function f defined on a closed interval (of finite length), Lebesgue's definition extends Riemann's in the sense that whenever the Riemann integral of f exists, the Lebesgue one exists as well and has the same value.[1] Similar theorem holds in $\mathbb{R}^n$.

The most famous example of a function having the Lebesgue integral but not the Riemann one is given by

$$f(x) = \begin{cases} 1 & \text{for } x \text{ rational,} \\ 0 & \text{for } x \text{ irrational.} \end{cases}$$

It is clear that $\int_0^1 f(x)\,\mathrm{d}x$ in the Riemann sense does not exist, since any inner rectangle has height 0 but any outer rectangle height 1. Later on, we will see that this integral is defined in the Lebesgue sense and equals 0.

A skeptic may regard this example as an irrelevant curiosity. It should be honestly admitted that for most functions on $\mathbb{R}^n$ one may encounter "in practice," in particular for all continuous functions, the Riemann integral works. Then, why bother with Lebesgue's more complicated definition? After all, practically all introductory courses prefer the Riemann integral because of its simplicity.

There are (at least) two main reasons for using the Lebesgue integral.

[1] The assumption that the interval is closed and bounded is not too restrictive, since a Riemann integral is naturally defined only on such intervals. These integrals are sometimes called *proper* to distinguish them from *improper* Riemann integrals, which are defined on open or unbounded intervals, as limits of proper integrals. Once we define Lebesgue integral, we will see that it may happen that an improper Riemann integral exists, while Lebesgue does not—for example, $\int_0^\infty \frac{\sin x}{x}\,\mathrm{d}x$.

- (Better properties) Even when working with specific functions, we often need to apply general theorems, and theorems about the Riemann integral are weaker and have much more complicated assumptions.

  A specific reason leading to Lebesgue's work was that the Riemann integral behaves badly under taking limits, one of the basic devices in dealing with functions. That is, if $f_1, f_2, \dots$ is a sequence of functions and f is their pointwise limit, i.e., $f(x) = \lim_{n\to\infty} f_n(x)$ for every x, then ideally one would like to have

  $$\lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x.$$

  For the Riemann integral this may fail even in seemingly innocent cases, e.g., all the f_n continuous and uniformly bounded on $[a, b]$. On the other hand, for the Lebesgue integral it holds in considerable generality, although some mild assumptions are needed.

- (Integration on general spaces) The Riemann integral is easy to define on $\mathbb{R}$, or on $\mathbb{R}^n$, but for functions on curved surfaces, for example, it becomes problematic. In contrast, the Lebesgue integral is defined on any *measure space*, i.e., whenever we specify a way of measuring "volumes" in an abstract sense.

  This abstract theory of integration was developed only many years after Lebesgue's work, and since the 1930s it serves as a foundation of modern probability theory.

Exercise 0.1. Find a sequence of Riemann integrable functions (in this exercise they need not be continuous) $f_1, f_2, \dots : [0, 1] \to [0, 1]$ such that the pointwise limit $f = \lim_{n\to\infty} f_n$ exists but is not Riemann integrable. ✠

Let us remark that there is an alternative path to building the Lebesgue integral and measure: the *Daniell integral*. This integral is defined axiomatically, and the measure is then obtained as a by-product. Daniell's way has some advantages over the Lebesgue approach, e.g., greater generality. However, since for applications in

computer science and combinatorics, understanding measure seems more important than fine points of integration, and since the advantages of the Daniell integral really show only in fairly advanced contexts, we prefer the more traditional way, constructing the measure first.

1. Measure

As was indicated above, we want to define measure before defining integral. We begin with measuring subsets of the real line $\mathbb{R}$. This case may look much simpler at first sight than measuring areas in the plane, but actually, the main issues are already present in this setting.

Our goal is thus to define a sensible notion of "length" for all "reasonable" subsets $A \subseteq \mathbb{R}$. Ideally, we would of course prefer to define length for *every* subset of $\mathbb{R}$, but it turns that this is not possible to do in a sensible way, for reasons discussed later.

1.1. The Outer Lebesgue Measure. First we define the *outer Lebesgue measure*: this is a function λ^* that assigns a nonnegative real number, or ∞, to *every* set $A \subseteq \mathbb{R}$. The definition uses countable covers of A by open intervals:

Definition 1.1. *The* **outer Lebesgue measure** *of a set $A \subseteq \mathbb{R}$ is defined as*

$$\lambda^*(A) = \inf\left\{ \sum_{i=1}^{\infty} \ell(I_i) : I_1, I_2, \dots \text{ open intervals, } A \subseteq \bigcup_{i=1}^{\infty} I_i \right\},$$

where $\ell(I)$ denotes the length of the interval I; for $I = (a, b)$, $a < b$, we have $\ell(I) = b - a$.

The values attained by λ^* include $+\infty$, and so it is convenient to say that the range of λ^* are the *extended real numbers*, denoted by $\overline{\mathbb{R}}$. As a set, $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$, and the arithmetic operations are defined as expected; e.g., $17 + \infty = \infty$. Some expressions, such as $\infty - \infty$, remain undefined.

Some good properties of λ^*. First let us check that λ^* measures the simplest sets, *intervals*, in the intended way. This is not that easy to prove, actually: one needs to use the fact that every bounded subset of $\mathbb{R}$ has a supremum, or some equivalent property, e.g., compactness of closed intervals $[a,b]$ (see Chapter 6, Section 3).

A terminological remark is in order here—an interval can be closed, half-closed, or open. It can also be finite (i.e., both endpoints are real numbers) or infinite (at least one of the endpoints is $+\infty$ or $-\infty$). We will repeatedly use that fact that λ^* is *monotone*, i.e., $A \subseteq B$ implies $\lambda^*(A) \leq \lambda^*(B)$. This follows easily from the definition.

Lemma 1.2. *If I is an interval, then $\lambda^*(I) = \ell(I)$.*

Proof. We will only prove the lemma for closed finite intervals $I = [a,b]$ and leave the rest to the reader as an exercise.

Choose a positive real ε. Since $I \subseteq (a-\varepsilon, b+\varepsilon)$, we have $\lambda^*([a,b]) \leq b-a+2\varepsilon$. The value of ε can be chosen arbitrarily small, and so $\lambda^*([a,b]) \leq b-a$.

To prove the other inequality, we need to show that whenever $\{I_1, I_2, \dots\}$ is a collection of open intervals covering the interval $[a,b]$, then $\sum \ell(I_i) \geq b-a$. The famous *Heine–Borel theorem* implies that there exists a *finite* subcollection of the intervals that also covers $[a,b]$ (this is the place where one needs to use some deep properties of $\mathbb{R}$). By deleting the unused intervals, the sum $\sum \ell(I_i)$ can only decrease. We can therefore suppose without loss of generality that $\{I_1, I_2, \dots\}$ is finite.

One of the intervals I_i contains the point a; let us call it (a_1, b_1). If $b_1 < b$, then some interval, say (a_2, b_2), contains b_1. We continue in a similar fashion until we arrive at an interval (a_k, b_k) containing b. We have

$$\sum \ell(I_i) \geq (b_1 - a_1) + (b_2 - b_1) + \dots + (b_k - b_{k-1}) = b_k - a_1 > b - a.$$

□

We strongly recommend doing the next exercise before reading further. It shows, in particular, that the set of all rational numbers

in $[0,1]$ has outer measure 0. This illustrates that the result proved above, $\lambda^*(I) = \ell(I)$, has to rely on some property that differentiates the reals from the rationals. To appreciate this issue, it may be good to remember that the discovery of irrational numbers in ancient Greece was a great surprise and one of the peak intellectual achievements of that time.

Exercise 1.3. Let $A \subset [0,1]$ be countable. Prove that $\lambda^*(A) = 0$. ✠

Next, we will check that if a set A is covered by finitely many, or even countably many, subsets, then the outer measure of A cannot be larger than the sum of the outer measures of the subsets. This property is called **countable subadditivity** of λ^*, and it immediately solves the previous exercise.

Lemma 1.4. *If $A_1, A_2, \ldots$ is a countable system of subsets of $\mathbb{R}$, then*

$$\lambda^*\Big(\bigcup_{i=1}^{\infty} A_i\Big) \le \sum_{i=1}^{\infty} \lambda^*(A_i).$$

Proof. The union of sufficiently small covers of the sets A_i is a sufficiently small cover of the set $\bigcup A_i$. More precisely, if $\{I_j^{(i)}\}_{j=1}^{\infty}$ is a cover of A_i such that $\sum_j \ell(I_j^{(i)}) \le \lambda^*(A_i) + \frac{\varepsilon}{2^i}$, $i = 1, 2, \ldots$, then

$$\lambda^*\Big(\bigcup A_i\Big) \le \sum_{i,j} \ell(I_j^{(i)}) \le \sum_i \lambda^*(A_i) + \varepsilon.$$

This is true for every $\varepsilon > 0$. □

The failure of additivity. Now one may ask, why not take the outer measure λ^* as the desired notion of "length"? The reason is that one would like to have length at least **finitely additive**, meaning that the length of the union of finitely many *disjoint* sets equals sum of their lengths. But for the outer measure λ^* this may fail!

Here is an example showing the failure of a stronger property, **countable additivity**.

Example 1.5. *There exist countably many sets $A_1, A_2, \ldots \subset \mathbb{R}$, mutually disjoint, such that*

$$\sum_{i=1}^{\infty} \lambda^*(A_i) \ne \lambda^*\Big(\bigcup_{i=1}^{\infty} A_i\Big).$$

Proof. Call two numbers $a, b \in [0,1]$ *equivalent* if $a - b$ is rational. Let $V \subset [0,1]$, a *Vitali set*, be a set containing precisely one number from each equivalence class.

Let us enumerate all the rational numbers in $[-1,1]$ in a sequence $q_1, q_2, q_3, \ldots$, and let $A_i := V + q_i$ (the translation of V by q_i).

The A_i are clearly disjoint and contained in $[-1,2]$, and some thought reveals that they together cover $[0,1]$. Hence $\lambda^*([0,1]) = 1 \leq \lambda^*(\bigcup_{i=1}^{\infty} A_i) \leq \lambda^*([-1,2]) = 3$.

Now λ^* is, by definition, translation-invariant, and so $\lambda^*(A_i) = \lambda^*(V)$ for every i. So if $\lambda^*(V) = 0$, then $\sum_{i=1}^{\infty} \lambda^*(A_i) = 0$, and if $\lambda^*(V) > 0$, then $\sum_{i=1}^{\infty} \lambda^*(A_i) = \infty$. Even without knowing which of these possibilities actually holds, we can say for sure that $\sum_{i=1}^{\infty} \lambda^*(A_i)$ cannot be between 1 and 3. □

From the failure of countable additivity one can also derive the failure of finite additivity.

Exercise 1.6. Suppose (falsely!) that λ^* is finitely additive. Show that it then must be countably additive as well. ✠

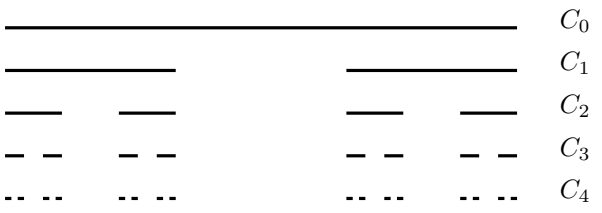
The outer Lebesgue measure in $\mathbb{R}^n$. For $\mathbb{R}^n$ instead of $\mathbb{R}$, intervals in the definition of λ^* are replaced with open axis-parallel boxes (in other words, Cartesian products of open intervals). Then it can be shown, with some more effort, that appropriate analogs of the statements above hold, i.e., boxes have the right outer measure and λ^* is countably subadditive.

The Banach–Tarski paradox. In dimension 3 and higher, the lack of finite additivity of the outer measure λ^* manifests itself in a particularly bizzare way. Namely, it can be shown, with fairly non-trivial proof, that a 3-dimensional ball B^3 can be decomposed into finitely many disjoint subsets $A_1, \ldots, A_k$ so that, after applying a suitable rotation and translation to each A_i, the resulting sets are again disjoint but they fill *two* disjoint copies of B^3. One says that B^3 and the two copies of B^3 are *equidecomposable*. In this particular case, it suffices to take $k = 5$.

Interestingly, a 2-dimensional analog of this paradox is impossible: two equidecomposable planar geometric figures must have the

same area. A standard treatment of the Banach–Tarski paradox and related problems is [**Wag93**], but many new results have been obtained since its publication, so readers interested in the current status of the various questions in [**Wag93**] should definitely consult the Internet.

Exercise 1.7. Let C be the Cantor "middle-third" set, which can be compactly defined as $\bigcap_{i=1}^{\infty} C_i$, where $C_0 = [0,1]$ and $C_i = \frac{1}{3}C_{i-1} \cup (\frac{1}{3}C_{i-1} + \frac{2}{3})$:



(a) Show that $\lambda^*(C) = 0$.

(b) Show that C is closed and uncountable. ✠

1.2. The Lebesgue Measure. The second step in the construction of the Lebesgue measure is defining a suitable system $\mathcal{E}$ of subsets of $\mathbb{R}$ (or $\mathbb{R}^n$, the definition is formally the same) such that the outer measure λ^* restricted to $\mathcal{E}$ becomes countably additive, and at the same time, $\mathcal{E}$ is as rich as possible. Then the Lebesgue measure λ we are after is the restriction of λ^* to this $\mathcal{E}$.

The complement of a set E is the set $\mathbb{R} \setminus E$. It will be denoted by E^c.

Definition 1.8. *We say that a set E is* **measurable** *if for every set $A \subseteq \mathbb{R}$ we have*

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E^c). \tag{1}$$

Let $\mathcal{E}$ be the system of all measurable subsets of $\mathbb{R}$. The **Lebesgue measure** *λ on $\mathbb{R}$ is the restriction of λ^* to $\mathcal{E}$.*

To check measurability of a set E, it suffices to verify the inequality $\geq$ in (1), since the opposite inequality is implied by the subadditivity of λ^*.

Exercise 1.9. Check that measurable sets can be approximated from the inside by compact sets. That is, for every $\varepsilon > 0$ and every measurable set E of finite measure there exists a compact set K such that $K \subseteq E$ and $\lambda^*(E \setminus K) < \varepsilon$. ✠

Likewise, measurable sets can be approximated from the outside by open sets. Using this, one can prove that a set E of finite outer measure is measurable if and only if for every $\varepsilon > 0$ there is a finite union F of open intervals such that $\lambda^*(F \Delta E) < \varepsilon$ (here $F \Delta E$ is the *symmetric difference* of the two sets, consisting of all points that belong to one of them but not the other one). That is, *every measurable set is nearly a finite union of intervals*, which is one of three informal observations known as *Littlewood's principles*.

We will now develop some properties of the system $\mathcal{E}$.

Observation 1.10. *If $\lambda^*(E) = 0$, then E is measurable.*

Proof. If A is an arbitrary set, then $A \cap E \subseteq E$ and $\lambda^*(A \cap E) \leq \lambda^*(E) = 0$. Similarly, $\lambda^*(A) \geq \lambda^*(A \cap E^c)$. Combining these two observations, we have

$$\lambda^*(A) \geq \lambda^*(A \cap E^c) = \lambda^*(A \cap E^c) + \lambda^*(A \cap E). \qquad \square$$

In the following exercise, diligent readers are asked to derive some basic properties of measurable sets. We recommend trying at least one of the parts, to gain some feeling for the definition of measurability.

Exercise 1.11. (a) Show that the interval $(0, \infty)$ is measurable.

(b) Prove that the complement of a measurable set is measurable.

(c) Verify that the union of two measurable sets is measurable, or more generally, that the union of countably many measurable sets is measurable. ✠

Set systems having the last two properties in the previous exercise are of fundamental importance in measure theory (and in probability theory), and they have a name:

Definition 1.12. *A* **σ-algebra** *(some authors also use the term* σ-field*) is a nonempty set system closed under complementation and countable unions.*

Combining the two properties in the definition, it is easy to see that σ-algebras are also closed under countable intersections.

Exercise 1.13. Let X be a finite set. Describe all σ-algebras on X. ✠

We have already seen one example of a σ-algebra—the measurable sets on $\mathbb{R}$ or $\mathbb{R}^n$. If $\mathcal{A}$ is a system of subsets of a set X, the intersection $\mathcal{F}$ of all σ-algebras containing $\mathcal{A}$ is also a σ-algebra. We call it *the smallest σ-algebra containing $\mathcal{A}$*. We also say that $\mathcal{F}$ is the σ-algebra *generated by* $\mathcal{A}$.

Borel sets. If $\mathcal{A}$ is the family of all open intervals on $\mathbb{R}$, then the elements of the smallest σ-algebra containing $\mathcal{A}$ are called *Borel sets*.

To define Borel sets, we only needed basic set operations (to define σ-algebra) and the notion of open sets. These tools are available in any metric or topological space, and Borel sets can be defined there as well, in exactly the same way.

The notion of Borel sets appears quite often. Basically, when one considers a space with measure *and* metric (or topology), one almost always assumes that all Borel sets are measurable.

The family of Borel sets in $\mathbb{R}$ is a σ-algebra, and so, together with open intervals, it contains all of their countable unions. Likewise, it contains all countable intersections of these countable unions. But even if we iterate countable intersections and countable unions a thousand times, there will still be some Borel sets that we have not created. Borel sets can be defined inductively in a manner similar to what we have just described, but in this case the induction needs to be *transfinite*—we need ω_1 steps, where ω_1 is the first uncountable ordinal number.

Exercise 1.14. Prove that every Borel set in $\mathbb{R}$ is Lebesgue measurable. ✠

The converse is far from true; see Exercise 1.18.

It remains to prove that the Lebesgue measure is countably additive.

Theorem 1.15. *If $E_1, E_2, \dots$ is a sequence of pairwise disjoint measurable sets, then*

$$\lambda\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \lambda(E_i).$$

Proof. Measurable sets form a σ-algebra, and so $\bigcup E_i$ is a measurable set.

We have already proved subadditivity, even for λ^*.

Let $E_1, E_2, \dots$ be disjoint measurable sets. We take $A = E_1 \cup E_2$ as the "testing" set in the definition of a measurable set. Since E_1 is measurable, this definition tells us that

$$\lambda(E_1 \cup E_2) = \lambda(E_1) + \lambda(E_2).$$

This, of course, can be extended to any finite family of disjoint measurable sets.

Let us now suppose that the sequence $E_1, E_2, \dots$ is countably infinite. Then

$$\lambda\Big(\bigcup_{i=1}^{\infty} E_i\Big) \geq \lambda\Big(\bigcup_{i=1}^{n} E_i\Big) = \sum_{i=1}^{n} \lambda(E_i)$$

for every natural number n, yielding $\lambda(\bigcup_{i=1}^{\infty} E_i) \geq \sum_{i=1}^{\infty} \lambda(E_i)$. □

On non-measurable sets. Probably the simplest known non-measurable set is the set V in Example 1.5. In order to define it, we needed to select one point from each of uncountably many equivalence classes, and this relies on the *axiom of choice* in set theory.

It is known that one cannot prove the existence of a non-measurable set without the axiom of choice. More precisely, the system of axioms of the usual (Zermelo–Fraenkel) set theory without the axiom of choice does not contradict the assumption "all subsets of $\mathbb{R}$ are Lebesgue measurable."

Since the axiom of choice is inherently nonconstructive, there is no way of fully describing a specific non-measurable set. You will never

see non-measurable sets in the daylight, but they may be lurking in the darkness—and you must protect your theorems against them.

Null sets and "almost everywhere." A set of measure zero is also often called a **null set**. One of the most often used properties of measure is that the union of finitely or countably many null sets is again a null set.

Here is a simple but typical way of using this. It is easy to check that a line in $\mathbb{R}^2$ has the (2-dimensional) Lebesgue measure 0. Hence the plane cannot be covered by countably many lines.

This gives us a very simple proof of existence of point sets in general position (no three collinear): having already constructed an n-point set P in general position, the points of P span finitely many lines, these do not cover $\mathbb{R}^2$, and so we can always add a new point to P while keeping general position. (Using transfinite induction, we can even obtain an uncountable set in general position in this way.) Similarly we can require no four of our points co-circular (since a circle has measure 0), and so on.

Here is another very common and useful piece of terminology: if Π is some property of a point in $\mathbb{R}^n$, and the set of points where Π does not hold is a null set, we say that Π holds **almost everywhere** (or for *almost all points*, etc.). Similarly, if $E \subset \mathbb{R}^n$ is a measurable set, we may say that Π holds almost everywhere on E. We will practice this terminology in the next remark.

A measurable set cannot be completely gray. Can we imagine what a measurable set looks like? One remarkable feature is that, on a sufficiently fine scale, it must be "grained"—it cannot be "gray" everywhere, i.e., to occupy half of each interval, say.

Let $E \subset \mathbb{R}^n$ be a measurable set. For a point $x \in \mathbb{R}^n$ we consider the limit

$$d_E(x) := \lim_{\delta \to 0} \frac{\lambda(B(x,\delta) \cap E)}{\lambda(B(x,\delta))},$$

where $B(x,r)$ denotes the (Euclidean) ball of radius r centered at x. The limit may not exist in general, but if it does, it is called the **density** of E at x.

The non-existence of a "completely gray" measurable set is expressed in the *Lebesgue density theorem*, asserting that every measurable $E \subseteq \mathbb{R}^n$ has density 1 at almost all of its points. That is, the set of those $x \in E$ for which $d_E(x)$ is undefined or smaller than 1 has Lebesgue measure 0.

Exercise 1.16. Prove that if $E \subseteq [0,1]$ is measurable with $\lambda(E) > 0$, then $E \cap V$ is not measurable, where V is as in Example 1.5. ✠

Exercise 1.17. (a) Consider a modified Cantor set (sometimes called the Smith–Volterra–Cantor set) $\tilde{C} = \bigcap_{i=1}^{\infty} \tilde{C}_i$, (cf. Exercise 1.7), where the deleted intervals gradually occupy a smaller and smaller proportion of the remaining part: We set $\tilde{C}_1 = [0,1]$ as before, and in the i-th step, $i = 1, 2, \ldots$, we have $\tilde{C}_i$ consisting of 2^{i-1} intervals of equal length. To obtain $\tilde{C}_{i+1}$, we remove the middle part of length 2^{-2i} from each interval in $\tilde{C}_i$. Show that $\tilde{C}$ is measurable and $\lambda(\tilde{C}) > 0$.

(b) Check that the characteristic function $\chi_{\tilde{C}}$ of $\tilde{C}$ is not Riemann integrable.

(c) Construct a sequence of continuous, and hence Riemann integrable, functions that converge to $\chi_{\tilde{C}}$. ✠

Exercise 1.18. (a) Construct a homeomorphism (a continuous map with continuous inverse) $\varphi\colon [0,1] \to [0,1]$ that maps the modified Cantor set $\tilde{C}$ from Exercise 1.17 to the "usual" Cantor set C (Exercise 1.7). Thus, homeomorphisms need not preserve null sets in general.

(b) Use (a) and Exercise 1.16 to exhibit a non-Borel measurable set. Hint: all subsets of the Cantor set are measurable. ✠

1.3. Measure Spaces and Important Examples. We started with $\mathbb{R}$, chose a particular family of subsets (which we called *measurable sets*), and defined a mapping from this family to extended real numbers (the Lebesgue measure). Albeit very important, this is just an instance of a more general concept.

Definition 1.19. *Let $\mathcal{F}$ be a σ-algebra. A* **measure** *on $\mathcal{F}$ is a function $\mu : \mathcal{F} \to \overline{\mathbb{R}}$ that is*

- *nonnegative,*
- *countably additive (if $E_1, E_2, \ldots$ are pairwise disjoint, then $\mu\left(\bigcup_j E_j\right) = \sum_j \mu(E_j)$), and*
- *assigns the value 0 to $\emptyset$.*

A **measure space** *is a triple $(X, \mathcal{F}, \mu)$, where X is a set, $\mathcal{F}$ is a σ-algebra on X, and μ is a measure on $\mathcal{F}$.*

As was the case with Lebesgue measure, $+\infty$ is an admissible value of μ. The elements of the σ-algebra $\mathcal{F}$ are called *measurable sets*. Usually the σ-algebra that we are referring to is clear from the context.

If we want to stress that we are talking about sets of real numbers satisfying Definition 1.8 (for which we have reserved the name measurable until now), we call them *Lebesgue measurable*.

If a property holds everywhere except for a set A with $\mu(A) = 0$, we say that it holds *μ-almost everywhere*.

Exercise 1.20. Let $A_1 \supseteq A_2 \supseteq \cdots$ be a nonincreasing sequence of μ-measurable sets with $\bigcap_{n=1}^{\infty} A_n = \emptyset$.

(a) Find an example, in which μ is the Lebesgue measure on $\mathbb{R}$, such that $\lim_{n\to\infty} \mu(A_n) \neq 0$.

(b) Now suppose that $\mu(A_1) < \infty$, and prove $\lim_{n\to\infty} \mu(A_n) = 0$. ✠

We proceed to important examples.

Counting and Dirac measures. If X is a set and $\mathcal{F}$ is the family of all subsets of X, the *counting measure* assigns to each set its number of elements. This measure does not distinguish infinite cardinalities, infinite sets are simply assigned $+\infty$.

Now let us fix a point $x_0 \in X$, and let $\mathcal{F}$ consist of all subsets of X again. The *Dirac measure* δ_{x_0} assigns 0 or 1 to each $A \in \mathcal{F}$, depending on whether $x_0 \in A$.

Hausdorff measure and Hausdorff dimension. The Lebesgue measure in $\mathbb{R}^n$ measures n-dimensional volume, but what if we want to measure the area of a surface in $\mathbb{R}^3$, for example?

A very general tool for measuring d-dimensional volumes in $\mathbb{R}^n$ is the d-dimensional **Hausdorff measure** H^d. Similar to the Lebesgue measure, the definition of H^d has two steps: first we define an outer measure H^{d*} on $\mathbb{R}^n$, and then we restrict it to an appropriate σ-algebra of measurable sets, where measurability is defined as in Definition 1.8, but with H^{d*} instead of λ^*.

The interesting part is the definition of H^{d*}, which can actually be used not only for subsets of $\mathbb{R}^n$, but for subsets of an arbitrary metric space.

First we define the **diameter** $\operatorname{diam} U$ of a set U as the supremum of distances of points $x, y \in U$. Next, for a set A and a real number $\delta > 0$, we set

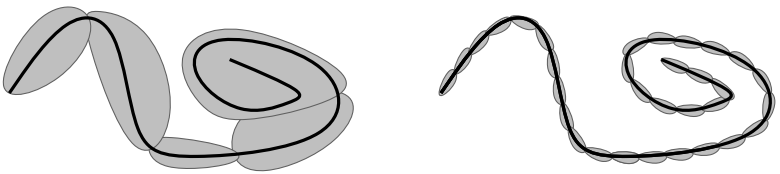
$$H^d_\delta(A) := \inf\Big\{ \sum_{i=1}^{\infty} (\operatorname{diam} U_i)^d : \bigcup_{i=1}^{\infty} U_i \supseteq A, \operatorname{diam} U_i < \delta \text{ for all } i \Big\}.$$

Finally,

$$H^{d*}(A) := \lim_{\delta \to 0} H^d_\delta(A)$$

(the limit, possibly infinite, exists since $H^d_\delta(A)$ is obviously nondecreasing as δ decreases).

To see how this works, let us first see how H^1 measures the length of a planar curve. For a given $\delta > 0$, we cover the curve with at most countably many sets of diameter smaller than δ, and we add up their diameters.



For some δ we may get much less than the length of the curve, but as δ gets smaller, the covering sets are forced to trace the curve more

and more closely, and it is at least intuitively plausible that we get a good notion of length.

Next, let us look at H^2 in the plane; then we are summing *squared* diameters. Let us consider the unit square Q, for example. Given $\delta > 0$, we can cover Q with $O(\delta^{-2})$ disks of diameter δ, and so $H^2_\delta(Q) = O(1)$.

On the other hand, we do need at least const $\cdot\, \delta^{-2}$ of sets of diameter at most δ in any covering: Choose an $n \times n$ square grid in Q. Then every two points are at least $\frac{1}{n}$ apart, and so each set of diameter less than $\frac{1}{n}$ can cover at most one grid point.

Hence $H^2(Q)$ is some positive constant. Actually, it turns out that in $\mathbb{R}^n$, the n-dimensional Hausdorff measure H^n is a constant multiple of the Lebesgue measure,[2] where the constant factor is the volume of a ball of radius $\frac{1}{2}$.

It may seem that the "most economic" covering should always be with balls of diameter δ. However, this is not the case in general: it can be shown that if we restrict only to coverings with balls, the measure we get is different from H^d for some sets.

An interesting feature is that d, the dimension, in the definition of H^d need not be an integer—the definition makes sense for every real $d \geq 0$. It is easy to check that $H^d \geq H^{d'}$ for $d \leq d'$. It turns out that for every set A there is exactly one "right" value d_0 of d, in the sense that for $d < d_0$ we have $H^d(A) = \infty$ and for $d > d_0$ we have $H^d(A) = 0$. (For d_0 itself, $H^{d_0}(A)$ can be a positive number, but also 0 or ∞.) This d_0 is called the **Hausdorff dimension** of A.

For example, the Cantor set C in Exercise 1.7 has Hausdorff dimension $\log_3 2 \approx 0.631$. Indeed, for every k, we can cover C by 2^k intervals of length 3^{-k}, and we cannot do much better. Most likely the reader has heard about *fractals*; the Cantor set is a very simple example of this popular concept. The Hausdorff dimension is very useful for measuring fractals, but also various exceptional sets in results of mathematical analysis, for example. (There are many other concepts of dimension, e.g., several topological ones, but these are usually only integer-valued.)

[2] Some authors re-normalize the Hausdorff measure so that H^n in $\mathbb{R}^n$ is the same as the Lebesgue measure.

Measuring the unit sphere. Quite often one needs a rotation-invariant measure μ on the unit sphere S^n in $\mathbb{R}^{n+1}$. The n-dimensional Hausdorff measure will do, but a more convenient and elementary definition is this: for a set $A \subseteq S^n$, define $\tilde{A}$ as the union of all segments connecting points of A to the center of the sphere, and define $\mu(A) := \lambda(\tilde{A})/\lambda(B^{n+1})$, where $B^{n+1} = \widetilde{S^n}$ is the unit ball.

The Haar measure. There are many ways of choosing a measure on an arbitrary set X. But a very useful fact of mathematics is that if X is a *group*, then, under reasonable assumptions, there is an essentially unique, canonical way of choosing a measure on X that is compatible with the group structure in a suitable sense. This result is known as *Haar's theorem*, and the respective measure as the *Haar measure*.

One thing that we take for granted when dealing with areas in the plane or volumes in $\mathbb{R}^n$ is that the volume of a set does not change if we translate the set. The Lebesgue measure satisfies this requirement—if the measure of a set $S \subset \mathbb{R}^n$ is defined, it is the same as the measure of the set $c+S$, for any $c \in \mathbb{R}^n$. We say that Lebesgue measure is invariant under the action of the group $\mathbb{R}^n$, with addition as the group operation.

Of course, there is nothing special about the additive group of reals (or $\mathbb{R}^n$). We would like to extend this concept to other (suitable) groups. This is the idea behind Haar measure. The Lebesgue measure on $\mathbb{R}^n$ is actually the completion of the Haar measure on $\mathbb{R}^n$ with the group operation of vector addition.

As another tiny example, let us consider the multiplicative group of positive real numbers. Clearly, Lebesgue measure is not invariant with respect to this group. But we can define a measure on the set of positive real numbers that does have this property, as follows:

$$\mu(S) := \int_S \frac{1}{x}\,\mathrm{d}x.$$

It is easy to check that whenever $0 < a < b$ and c is a positive real number, we indeed have $\mu([a,b]) = \mu([ca,cb])$.

One of the most important types of groups in all mathematics (and physics) are *matrix groups*; these are subgroups of the group $\mathrm{GL}(n,\mathbb{K})$ of all $n \times n$ invertible matrices over a field $\mathbb{K}$.

Let us consider the case $\mathbb{K} = \mathbb{R}$. Then $\mathrm{GL}(n,\mathbb{R})$ can be regarded as a subset of $\mathbb{R}^{n^2}$, which defines a metric (and topology) on $\mathrm{GL}(n,\mathbb{R})$, and allows us to speak of Borel sets in $\mathrm{GL}(n,\mathbb{R})$.

In contrast to the simple cases of additive or multiplicative groups of reals, we now need to start distinguishing left- and right-invariance. Haar's theorem implies that there is a measure ν on $\mathrm{GL}(n,\mathbb{R})$ that is *left translation-invariant* under matrix multiplication; that is, for every Borel set $S \subseteq \mathrm{GL}(n,\mathbb{R})$ and every matrix $A \in \mathrm{GL}(n,\mathbb{R})$, we have

$$\nu(AS) = \nu(S), \text{ where } AS = \{AB : B \in S\}.$$

Moreover, if we require certain reasonable properties of ν (finiteness on compact sets and regularity; see below), then ν is a *unique* left translation-invariant Borel measure on $\mathrm{GL}(n,\mathbb{R})$—more precisely, unique up to a scalar multiple (since if ν works, then 10ν works, too ...).

This ν is also right translation-invariant, $\nu(SA) = \nu(S)$; this is a special property of $\mathrm{GL}(n,\mathbb{R})$, not shared by all groups covered by Haar's theorem.

Let us now consider the subgroup $\mathrm{SO}(n,\mathbb{R}) \subseteq \mathrm{GL}(n,\mathbb{R})$ of all rotations in $\mathbb{R}^n$ (orthogonal matrices with determinant 1, that is). Then, again, there is a unique, up to scalar multiple, reasonable left translation-invariant Borel measure ν' on $\mathrm{SO}(n,\mathbb{R})$. It again happens to be right translation-invariant as well. Let us also note that ν' is *not* the restriction of the ν from above (since it can be shown that $\nu(\mathrm{SO}(n,\mathbb{R})) = 0$). While $\nu(\mathrm{GL}(n,\mathbb{R})) = \infty$, we have $\nu'(\mathrm{SO}(n,\mathbb{R}))$ finite.

Let us mention in passing that this ν' can also be used to define a rotation-invariant measure on the sphere: for a Borel set $E \subseteq S^{n-1}$, we set $\mu(E) := \nu'\{A \in \mathrm{SO}(n,\mathbb{R}) : Ax_0 \in E\}$, where $x_0 \in S^{n-1}$ is a fixed point. It can be shown that, up to scalar multiple, this is the same measure on S^{n-1} as the one we have introduced earlier in a more pedestrian way.

Stating Haar's theorem in general requires introducing several notions. This part is more technical than the rest of the chapter and a less experienced reader may skip it. The main message is that nice enough groups possess essentially unique reasonable left translation-invariant measure.

Haar's theorem is concerned with the situation where the group G in question is a **topological group**. This means that G is also a topological space[3] such that the group operations, multiplication and inversion, are continuous. Moreover, we want G to be a *locally compact topological space*, which in the case of a topological group means that the unit element $e \in G$ has a compact neighborhood (in the metric case, a sufficiently small closed ball around e is compact). We also need that G be *Hausdorff*, which is a fairly standard condition which holds for all metric spaces, for example, and only very exotic topological spaces fail to satisfy it. All of these conditions are satisfied for the matrix groups mentioned above.

The definition of a **regular measure** has two parts. A measure μ is *outer-regular on Borel sets* if $\mu(E) = \inf\{\mu(U) : U \supseteq E, U \text{ open}\}$ for every Borel set E (i.e., Borel sets can be approximated from outside by open sets), while μ is *inner-regular on open sets* if $\mu(U) = \sup\{\mu(K) : K \subseteq U, K \text{ compact}\}$ for every open set U. A *regular* measure is one that is both outer-regular on Borel sets and inner-regular on open sets.

Now we can state Haar's theorem (without proof):

Theorem 1.21 (Haar's theorem)**.** *Let G be a Hausdorff and locally compact topological group. Then there is a measure μ on the σ-algebra of Borel sets in G, unique up to a multiplicative constant, with the following properties:*

- *μ is nontrivial (attains some values in $(0, \infty)$);*
- *μ is left translation-invariant;*
- *μ is finite on all compact subsets of G;*
- *μ is regular in the sense above.*

[3] Readers not familiar with topology can think of a metric space, or a subspace of some $\mathbb{R}^n$ or possibly look into Chapter 6.

A μ as in the theorem is called the **left Haar measure**. One can define right Haar measure analogously. For compact G or commutative G, and in many other cases of interest, the left and right Haar measures coincide. However, for example for the group of all *invertible affine transformations of* $\mathbb{R}$, of the form $x \mapsto ax + b$, $a \neq 0$, with the operation of composition, the left and right Haar measures are not the same.

For the usual matrix groups, explicit formulas for the Haar measure are known, but often one needs just the existence (and sometimes uniqueness).

In addition to the examples already given, we mention that the Haar measure on $\mathrm{GL}(n, \mathbb{R})$ is given by

$$\nu(E) = \int_E \frac{1}{|\det(X)|^n} \, \mathrm{d}X$$

(Lebesgue integral in $\mathbb{R}^{n^2}$; see below).

2. The Lebesgue Integral

2.1. Measurable Functions. We are getting close to defining the Lebesgue integral of a function f. In many contexts (probability theory etc.), it is useful to do it in a general measure space $(X, \mathcal{F}, \mu)$, not just in $\mathbb{R}$. Unfortunately, we cannot define the integral for *all* functions, but only for *measurable* ones. Measurable functions preserve the structure of measure spaces, much like continuous functions preserve the structure of topological spaces (the preimage of an open set is an open set).

Definition 2.1. *Let X and Y be sets equipped with σ-algebras $\mathcal{F}$ and $\mathcal{G}$, respectively. A function $f\colon X \to Y$ is* **measurable** *if $f^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{G}$.*

The next exercise provides a useful criterion of measurability.

Exercise 2.2. If $\mathcal{G}$ in Definition 2.1 is the smallest σ-algebra containing a set system $\mathcal{H}$, then, to show that f is measurable, one only needs to verify that $f^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{H}$. Prove that. ✠

When both $X = Y = \mathbb{R}$, the usual convention (somewhat asymmetric!) is that $\mathcal{F}$ is the σ-algebra of Lebesgue measurable sets as defined in Section 1.2 and $\mathcal{G}$ are the Borel sets. Using Exercise 2.2, we can rewrite the definition as follows.

Definition 2.3. *Let $D \in \mathcal{F}$. A function $f\colon D \to \overline{\mathbb{R}}$ is* **measurable** *if the set $\{x \in D : f(x) \in I\}$ is measurable for every interval $I \subseteq \overline{\mathbb{R}}$.*

In Section 1.2 we mentioned the first of three *Littlewood's principles*. The other two have to do with measurable functions. They say that every measurable function is nearly continuous, and that every convergent sequence of measurable functions is nearly uniformly convergent. We will not formalize these observations here, but even in this form they are useful to provide some intuition about measurable sets and functions.

For a real function f we define its **positive part** f^+ as $\max\{f, 0\}$ and its **negative part** f^- as $\max\{-f, 0\}$. Besides the usual conventions regarding adding infinity to a real number and multiplying infinity by a real number, we define

$$0 \cdot (\pm\infty) = 0.$$

The expressions $\infty - \infty$, $\pm\infty/0$, $a/0$ (for real a) and $\pm\infty/\pm\infty$ remain undefined.

The following properties of measurable functions are easy to verify. If $f, f_1, f_2, \ldots$ are measurable extended-real valued functions on $D \in \mathcal{F}$, then $|f|$, f^+, f^-, $\sup f_j$, $\inf f_j$, $\limsup f_j$, $\liminf f_j$ are measurable on D, $1/f$ is measurable on $\{x \in D; f(x) \neq 0\}$, and $f_1 + f_2$, $f_1 - f_2$, $f_1 f_2$, f_1/f_2 are measurable wherever these expressions are defined. Moreover, the set D' of points where $\lim f_j$ exists is measurable, and $f = \lim f_j$ is measurable on D'.

In the definition of the Lebesgue integral, *simple functions* play a role similar to the role of divisions of an interval in case of the Riemann integral.

Definition 2.4. *Let X be a set, $\mathcal{F}$ a family of subsets of X, and $D \in \mathcal{F}$. A function $f\colon D \to \mathbb{R}$ is* **simple** *if it is a linear combination of characteristic functions of sets in $\mathcal{F}$.*

The term *linear combination* implies that the sum is finite and the coefficients are real (no infinities). The definition makes sense for any set system $\mathcal{F}$, but the most important instance is when $\mathcal{F}$ is a σ-algebra. In this case a function is simple if and only if it is measurable and attains a finite number of values, all of them finite.

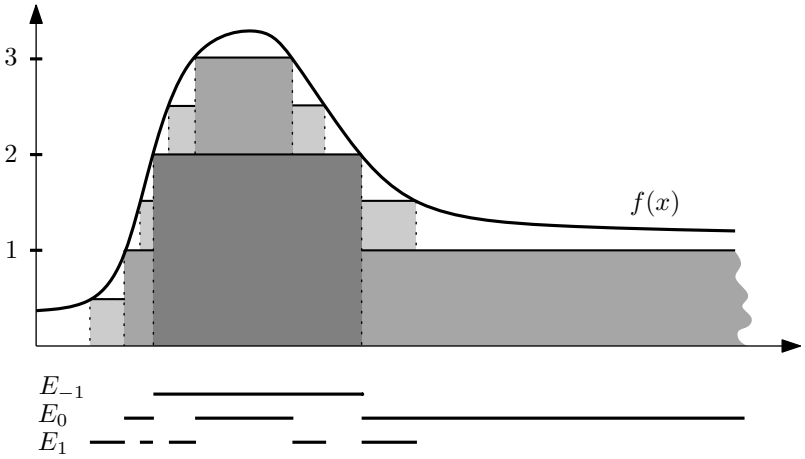
Proposition 2.5. *Let X be a set, $\mathcal{F}$ a σ-algebra on X, and f a nonnegative measurable function on $D \in \mathcal{F}$. Then one can find simple functions $f_1, f_2, \ldots$ that converge pointwise to f and such that $f_k \leq f_{k+1}$ for all k.*

Proof. We present a proof assuming that f attains only finite values. The extension so that values $+\infty$ are handled as well is not difficult and we leave it to the reader.

For $j \in \mathbb{Z}$ let us define $P_j = \bigcup_{i \text{ odd}} [i2^{-j}, (i+1)2^{-j})$ and $E_j = \{x \in D : f(x) \in P_j\}$. In other words, $x \in E_j$ if and only if 1 occupies the j-th position of the binary representation of $f(x)$. Since the P_j are Borel, the E_j are measurable. It is easy to see that

$$f = \sum_{j=-\infty}^{\infty} 2^{-j} \chi_{E_j}.$$

The functions $f_k = \sum_{j=-k}^{k} 2^{-j} \chi_{E_j}$ are simple and converge to f from below. Here is an illustration of this construction:



□

2.2. Lebesgue Integral, Convergence Theorems. Let $(X, \mathcal{F}, \mu)$ be a measure space. The Lebesgue integral of a function f on a set $D \in \mathcal{F}$ is usually denoted by $\int_D f \, \mathrm{d}\mu$. One can also encounter the notation $\int_D f$, $\int_D f(x) \, \mathrm{d}\mu(x)$, or even $\int_a^b f(x) \, \mathrm{d}x$ if D is the interval $[a, b]$.

In the last mentioned case, the notation is the same as the usual notation for the Riemann integral. However, it turns out that every Riemann integrable function on a closed interval is measurable, and that its Riemann and Lebesgue integrals coincide.

We will define the Lebesgue integral in three steps: for simple functions, then for nonnegative measurable functions, and finally for measurable functions with arbitrary signs.

Definition 2.6. *Let $(X, \mathcal{F}, \mu)$ be a measure space and $f = \sum_i a_i \chi_{A_i}$ a simple function on $D \in \mathcal{F}$. We define*

$$\int_D f d\mu := \sum_i a_i \mu(A_i).$$

There are many ways to write a given simple function. For instance, $\chi_{[0,2]} + \chi_{[1,3]}$ is the same function as $\chi_{[0,1]} + 2 \cdot \chi_{[1,2]} + \chi_{[2,3]}$. It turns out that the value of the integral does not depend on the way we write the function, but we skip the proof.

Definition 2.7. *Let $(X, \mathcal{F}, \mu)$ be a measure space and $D \in \mathcal{F}$.*

- *If f is a nonnegative measurable function on D, then*

$$\int_D f \, \mathrm{d}\mu = \sup \left\{ \int_D \psi \, \mathrm{d}\mu : \psi \text{ is a simple function with } \psi \leq f \right\}.$$

- *If f is a measurable function on D, then $\int_D f = \int_D f^+ - \int_D f^-$, provided the difference makes sense (at least one of the integrals on the right side is finite). If the difference is $\infty - \infty$, the integral is not defined.*

Perhaps slightly confusingly, we call f *integrable* whenever the integral $\int_D f \, \mathrm{d}\mu$ not only exists, but, moreover, is finite.

Now the reason for considering only measurable functions becomes more clear. Recall that a function is Riemann integrable if the supremum of the vertical rectangles that fit under the curve is

equal to the infimum of the rectangles that approximate the curve from above. For the Lebesgue integral the situation is similar: for nonnegative functions f, we would like to see the equality

$$\inf_{f\le\psi}\int\psi=\sup_{\varphi\le f}\int\varphi$$

where ψ and φ are simple functions. If f is a bounded function on a set of finite measure, then the above equality is true if and only if there is a measurable function g that is equal to f almost everywhere.

Let us gather some basic properties of the Lebesgue integral. While the proofs are not difficult, proving at least some of them might help the reader to gain some insight and practice the definitions.

Proposition 2.8. *Let $D\in\mathcal{F}$ and let f,g be measurable functions on D.*

(i) *If f is nonnegative, $D_1,D_2\in\mathcal{F}$, and $D_1\subseteq D_2\subseteq D$, then $\int_{D_1} f\,\mathrm{d}\mu\le\int_{D_2} f\,\mathrm{d}\mu$.*

(ii) *If $D_1,D_2\in\mathcal{F}$, $D_1\cap D_2=\emptyset$, and $D_1\cup D_2=D$, then $\int_D f\,\mathrm{d}\mu=\int_{D_1} f\,\mathrm{d}\mu+\int_{D_2} f\,\mathrm{d}\mu$.*

(iii) *If $\int_D|f|\,\mathrm{d}\mu<\infty$, then $|f|<\infty$ almost everywhere.*

(iv) *If $\int_D|f|\,\mathrm{d}\mu=0$, then $f=0$ almost everywhere.*

(v) *If f,g have integrals and $f\le g$ almost everywhere, then $\int_D f\,\mathrm{d}\mu\le\int_D g\,\mathrm{d}\mu$.*

(vi) *If $\int_D g\,\mathrm{d}\mu<\infty$ and $|f|\le g$ almost everywhere, then f is integrable.*

Exercise 2.9. Prove that if φ and ψ are nonnegative simple functions on $D\in\mathcal{F}$, then

$$\int_D\varphi+\psi=\int_D\varphi+\int_D\psi.$$ ✠

The Lebesgue integral is defined very generally and it is useful in a broad spectrum of situations. We do not impose many restrictions on the σ-algebra and the measure that we are using, but we pay for this by some loss of intuition. For instance, we saw in Exercise 2.9 that linearity of the integral is more or less obvious for simple functions. Linearity holds for general measurable functions as well,

but the proof is far from trivial. The easiest approach is probably through a monotone convergence theorem, as we will now show. We begin with another result, traditionally called a lemma.

Lemma 2.10 (Fatou's lemma)**.** *Let $D \in \mathcal{F}$ and let $\{f_n\}_{n=1}^{\infty}$ be a sequence of nonnegative measurable functions defined on D, converging to a function f. Then*

$$\int_D f \leq \liminf_{n\to\infty} \int_D f_n.$$

Proof. It is enough to show that whenever φ is a nonnegative simple function satisfying $\varphi \leq f$, then $\int_D \varphi \, \mathrm{d}\mu \leq \liminf \int_D f_n \, \mathrm{d}\mu$.

If $\int_D \varphi \, \mathrm{d}\mu < \infty$, then the set $A = \{x \in D : \varphi(x) > 0\}$ is measurable and of finite measure. We choose an $\varepsilon > 0$ and let

$$A_n = \{x \in D : \text{ for all } k \geq n \text{ we have } f_k(x) \geq (1-\varepsilon)\varphi(x)\}.$$

The sequence $\{A_n\}$ is increasing, i.e., $A_n \subseteq A_{n+1}$ for all n, and $A \subseteq \bigcup A_n$. It follows that $\lim_{n\to\infty} \mu(A \setminus A_n) = 0$ (see Exercise 1.20), and there is an n such that $\mu(A \setminus A_k) < \varepsilon$ for all $k \geq n$. If M is the maximum of the function φ, then for $k \geq n$ we have

$$\begin{aligned} \int_D f_k &\geq \int_{A_k} f_k \geq (1-\varepsilon) \int_{A_k} \varphi = (1-\varepsilon) \int_D \varphi - (1-\varepsilon) \int_{A\setminus A_k} \varphi \\ &\geq (1-\varepsilon) \int_D \varphi - \int_{A\setminus A_k} \varphi \geq (1-\varepsilon) \int_D \varphi - \varepsilon M. \end{aligned}$$

But then $\liminf \int_D f_k \geq (1-\varepsilon) \int_D \varphi - \varepsilon M$. The choice of $\varepsilon > 0$ was arbitrary, and so $\liminf \int_D f_k \geq \int_D \varphi$.

The situation is similar when $\int_D \varphi = \infty$. In this case, we can find a number $a > 0$ such that the set $A = \{x \in D : \varphi(x) > a\}$ has an infinite measure. We set

$$A_n = \{x \in D : \text{ for all } k \geq n \text{ we have } f_k(x) \geq a\}.$$

The sequence A_n is again increasing, and its union covers A. A variation on the theme of Exercise 1.20, which we leave to the reader, shows that $\lim \mu(A_n) = \mu(A) = \infty$. Since $\int_D f_n \geq a\mu(A_n)$, we also have $\lim \int_D f_n = \infty = \int_D \varphi$. □

The conclusion of Fatou's lemma also holds with the weaker assumption that the f_n converge to f almost everywhere. The proof is almost identical—we just need to separate the set (of measure zero) on which the functions do not converge.

Here is one of the promised results stating that integration and limit can be interchanged under suitable conditions.

Theorem 2.11 (Monotone convergence theorem)**.** *Let $D \in \mathcal{F}$ and let $\{f_n\}$ be a sequence of nonnegative functions that converge to a function f almost everywhere on D. If, moreover, $f_n \le f$ for all n, then*

$$\int_D f = \lim \int_D f_n.$$

Proof. Since $f_n \le f$, we also have $\int_D f_n \le \int_D f$. It follows that

$$\int_D f \le \liminf \int_D f_n \le \limsup \int_D f_n \le \int_D f. \qquad \square$$

Now, finally, we can prove the linearity of the Lebesgue integral. Lemma 2.10 and Theorem 2.11 are important in their own right, and this is only one of their many applications.

Theorem 2.12. *If f and g are integrable measurable functions on $D \in \mathcal{F}$ and $c_1, c_2 \in \mathbb{R}$, then*

$$\int_D (c_1 f + c_2 g) = c_1 \int_D f + c_2 \int_D g.$$

Proof. It easily follows from the definition of integral that $\int_D cf = c \int_D f$.

Let f and g be nonnegative measurable functions. If $\{\varphi_n\}$ and $\{\psi_n\}$ are sequences of simple functions converging to f and g, respectively, then $\{\varphi_n + \psi_n\}$ is a sequence of simple functions converging to $f + g$. Assuming, as we may, $\varphi_n \le f$ and $\psi_n \le g$ for all n and using Theorem 2.11 and Exercise 2.9, we get

$$\int_D (f+g) = \int_D \lim(\varphi_n + \psi_n) = \lim \left(\int_D \varphi_n + \int_D \psi_n \right) = \int_D f + \int_D g.$$

If f and g are general measurable functions, we partition D into sets where both are nonnegative, both negative, etc., and we treat each of these sets separately. $\square$

The following theorem, due to Lebesgue, is another basic and very useful tool.

Theorem 2.13 (Dominated convergence theorem)**.** *Suppose that $D \in \mathcal{F}$, g is an integrable function on D, and $\{f_n\}$ is a sequence of measurable functions such that $|f_n| \le g$ on D and the f_n converge to f almost everywhere on D. Then*

$$\int_D f = \lim \int_D f_n.$$

Proof. The functions $g - f_n$ are nonnegative, so we can apply Fatou's lemma and obtain

$$\int_D (g - f) \le \liminf \int_D (g - f_n).$$

Since $|f| \le g$, the function f is integrable and we have

$$\int_D g - \int_D f \le \int_D g - \limsup \int_D f_n.$$

It follows that $\int_D f \ge \limsup \int_D f_n$. Starting with $g + f_n$ in place of $g - f_n$, we conclude that $\int_D f \le \liminf \int_D f_n$. □

2.3. Product Measures, Fubini's Theorem. In combinatorics and counting, one of the most useful tricks is to interchange the order of summation, $\sum_i \sum_j = \sum_j \sum_i$. Fubini's theorem, treated in this section, provides an analogous tool for Lebesgue integration. For example, given a reasonable function $f\colon \mathbb{R}^2 \to [0, \infty)$, Fubini's theorem allows us to compute the volume between the graph of f and the xy plane by double integration: we can first integrate f with respect to the variable x, treating y as a constant, and then integrate the result with respect to y. Alternatively, we can first integrate with respect to y and then with respect to x.

As we will see, Fubini's theorem has several somewhat subtle assumptions. But as usual, it is much simpler and more general than an analogous result for the Riemann integral. We begin with some notions appearing in Fubini's theorem.

Definition 2.14. *Let $(X, \mathcal{F}, \mu)$ be a measure space. We say that μ is* **σ-finite** *if there is a sequence $X_1, X_2, \ldots$ of sets in $\mathcal{F}$ such that $\mu(X_j) < \infty$ for all j and $X = \bigcup_{j=1}^{\infty} X_j$.*

We say that μ is **complete** *if every subset of a set of measure zero is measurable.*

These two requirements ensure that our measure space is not overly exotic. For instance, $\mathbb{R}^n$ with the Lebesgue measure satisfies both of them.

Let $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{G}, \nu)$ be measure spaces such that the measures μ and ν are σ-finite. The elements of $\mathcal{F} \times \mathcal{G}$ are called *measurable rectangles*. For $A \in \mathcal{F}$ and $B \in \mathcal{G}$ we define

$$(\mu \times \nu)(A \times B) := \mu(A)\nu(B).$$

Then the **product measure** $\mu \otimes \nu$ is constructed in two steps: we first construct the corresponding outer measure $(\mu \times \nu)^*$, and then we define the σ-algebra $\overline{\mathcal{H}}$ of measurable sets with respect to $(\mu \times \nu)^*$, analogous to the way used for the Lebesgue measure (Definition 1.8). Then $\mu \otimes \nu$ is the restriction of $(\mu \times \nu)^*$ to $\overline{\mathcal{H}}$.

It can be shown that $\mu \otimes \nu$ is complete and coincides with $\mu \times \nu$ on all measurable rectangles, that $\overline{\mathcal{H}}$ is the smallest σ-algebra supporting a measure with these two properties, and that if both μ and ν are σ-finite, then $\mu \otimes \nu$ provides the only possible extension of $\mu \times \nu$ to the σ-algebra generated by the measurable rectangles.

We also remark that in many sources, the definition of the product measure is different: it lives on the σ-algebra generated by measurable rectangles, and it need not be complete. However, in this way, the product of the Lebesgue measure on $\mathbb{R}$ with itself is not the Lebesgue measure on $\mathbb{R}^2$, and so we prefer the definition above (which does guarantee that the $(n+m)$-dimensional Lebesgue measure is the product of the n- and m-dimensional Lebesgue measures).

Here is the main theorem of this section, which we state without proof.

Theorem 2.15 (Fubini's theorem)**.** *Suppose that* $(X, \mathcal{F}, \mu)$ *and* $(Y, \mathcal{G}, \nu)$ *are measure spaces with the measures* μ *and* ν *complete and* σ*-finite, and let* $(X \times Y, \mathcal{U}, \mu \otimes \nu)$ *be their (complete) product. Let* f *be a* $\mathcal{U}$*-measurable and integrable function on* $X \times Y$*. Then the following hold.*

(i) *For* μ*-almost all* x*, the function* f_x *defined by* $f_x(y) := f(x, y)$ *is an integrable function on* Y*. An analogous statement holds for* $f_y(x) := f(x, y)$*.*

(ii) $\int_Y f(x, y)\,\mathrm{d}\nu(y)$ *is an integrable function on* X*, and* $\int_X f(x, y)\,\mathrm{d}\mu(x)$ *is integrable on* Y*.*

(iii) $\int_X \left(\int_Y f\,\mathrm{d}\nu\right)\,\mathrm{d}\mu = \int_{X\times Y} f\,\mathrm{d}(\mu \otimes \nu) = \int_Y \left(\int_X f\,\mathrm{d}\mu\right)\,\mathrm{d}\nu$.

The same conclusion holds if we replace the requirement of integrability by the requirement that the function be nonnegative (this is known as *Tonelli's theorem*).

3. Foundations of Probability Theory

We assume that the reader has some background in probability; after all, it definitely belongs among the most important tools in other branches of mathematics and in science.

However, in order to introduce the basic notions of probability theory in a way that is both mathematically rigorous and sufficiently general, one needs the notion of measure—since probability *is* a kind of measure. Since measure is a relatively sophisticated notion, not so easily accessible to beginners, introductory courses on probability generally avoid laying proper foundations. Instead, they typically do finite probability spaces properly and geometric probability not at all, or in a semi-formal way.

Thus, having defined measure spaces, we want to point out their role in probability theory and state the usual axioms of probability.

3.1. Probability Spaces. What is a random point in the unit square? Everyone has some intuition about that; for example, imagine a square garden table, which is dry, but then it starts raining—the first drop hits the square at a random point.

In this case we are talking about the *uniform distribution* in the unit square, where no points are preferred to any others. It is tempting to say that every point should have the same probability, but this does not make much sense: there are infinitely many points in the square, and so the probability of hitting each particular one must be zero.

A reasonable thing to say is that for every geometric figure A in the square, the probability of hitting it should equal its area. So for geometric probability theory, we certainly need a definition of area. Moreover, if we want to conclude, for example, that hitting a point in the unit square with rational coordinates has zero probability, we are naturally led to countable additivity and thus to the notion of measure. However, the connection of contemporary probability theory to measure theory is much closer than just via geometric examples.

Axioms of probability. The current mathematical notion of probability took several hundred years to crystallize. It does not try to answer difficult philosophical questions like "What is randomness?", "Where does it come from?", "What is the meaning of probability in the real world?", etc., but it offers a mathematical model, which proves extremely successful in modeling real-world phenomena.

In most of the contemporary mathematical treatments of probability, the basic notion is a **probability space**, which is a mathematical model of some random process or experiment. As running examples, let us consider two simple experiments: three successive tosses of a fair coin (Example 3C), and picking a random point in the unit square (Example Sq).

A probability space is a triple $(\Omega, \mathcal{F}, \mathrm{P})$. The first component Ω is sometimes called the *sample space*, and it is a set consisting of all possible outcomes of the experiment. Each element $\omega \in \Omega$ is called an **elementary event**. For Example 3C, $\Omega_{3\mathrm{C}}$ consists of all possible three-letter sequences with letters H (heads) and T (tails): $\Omega = \{HHH, HHT, \ldots, TTT\}$. For Example Sq, $\Omega_{\mathrm{Sq}} = [0,1]^2$ consists of all points of the unit square.

The second component $\mathcal{F}$ of a probability space is a system of subsets of Ω. Each set $E \in \mathcal{F}$ is called an **event**. In Example 3C,

we admit *every* possible subset of Ω as an event, so $\mathcal{F}_{3C} = 2^{\Omega_{3C}}$. A concrete example of an event is "odd number of tails", which is the set $E = \{HHT, HTH, THH, TTT\}$. In Example Sq, the usual choice of $\mathcal{F}_{Sq}$ is the system of all Lebesgue measurable subsets of $[0,1]^2$: in this way, all reasonable geometric figures are events, and so we can talk about the probability of hitting the left half of the square, or hitting at most 0.1 from the center, etc., but "unreasonable" subsets of the square, i.e., sets that are not Lebesgue measurable, are not events.

The last component P of a probability space is a function that assigns a real number $\mathrm{P}(E)$, called the **probability of E**,[4] to every event $E \in \mathcal{F}$. In Example 3C, we consider all of the elementary events equally likely, and the probability of an event E is defined as $\frac{|E|}{|\Omega_{3C}|}$. In Example Sq, we set $\mathrm{P}_{Sq}(E) = \lambda(E)$, the Lebesgue measure. (This works because the whole unit square has measure 1; if we considered a random point in some other geometric figure Ω in $\mathbb{R}^n$, we would need to take the ratio $\frac{\lambda(E)}{\lambda(\Omega)}$.)

The triple $(\Omega, \mathcal{F}, \mathrm{P})$ should satisfy the following axioms, which were first presented in this form by Kolmogorov in the 1930s:

Kolmogorov's axioms for probability space $(\Omega, \mathcal{F}, \mathrm{P})$

(0) The system $\mathcal{F}$ of events forms a σ-algebra. Explicitly, this means that Ω is an event, the complement of an event is an event, and so is a countable union of events.

(1) $\mathrm{P}(E) \geq 0$ for every $E \in \mathcal{F}$.

(2) $\mathrm{P}(\Omega) = 1$ (the experiment always has some outcome).

(3) P is countably additive: whenever $E_1, E_2, \ldots$ is a sequence of mutually disjoint events, then

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mathrm{P}(E_i).$$

[4] In later chapters we will also be using $\mathrm{Prob}[E]$.

Axioms (1) and (2) are very intuitive, and so is finite additivity of probability. Countable additivity cannot really be substantiated by experience or intuition, but it becomes hard to avoid once we want to have reasonable notions of convergence of a sequence of random variables, which is a basic tool in probability theory and statistics.

A number of other natural properties of probability can be derived from the axioms, such as $\mathrm{P}(\emptyset) = 0$ or $\mathrm{P}(E) \leq \mathrm{P}(F)$ for $E \subseteq F$.

Comparing these axioms with the definition of a measure space, we find only one difference: the probability P is a measure with the additional condition $\mathrm{P}(\Omega) = 1$. Such a measure is called a **probability measure**.

A probability measure on a set Ω is also often referred to as a *probability distribution* on Ω.

Exercise 3.1. What is a probability space suitable for modeling the experiment "choosing three points a, b, c in the unit square, each of them uniformly distributed, the choices being mutually independent"? Hint: the triple needs to be represented by a single point in an appropriate higher-dimensional cube. ✠

Random real number and Benford's law. One mathematically simple but perhaps somewhat important remark is that there is no probability distribution on $\mathbb{R}$ in which "all real numbers have the same probability." If we wanted to make this vague phrase precise, we should require that the probability of hitting any interval I be proportional to its length. But any measure on $\mathbb{R}$ with this property must either be zero everywhere or assign measure ∞ to the whole $\mathbb{R}$, and so it cannot be a probability measure.

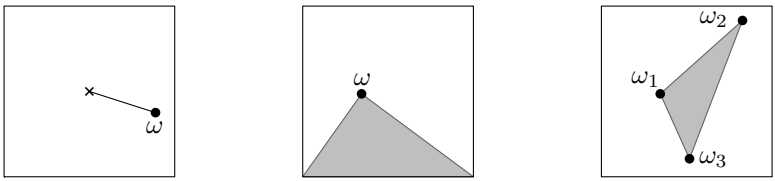
Yet in everyday life, in science, and at many other occasions we are confronted with various numbers that look quite random—your electricity bill, the area of the largest lake in your country, the number of followers of some Facebook group, etc.

We cannot resist mentioning a curious observation concerning such numbers. Namely, if we look at the leading digit of such "random" numbers, we encounter smaller digits considerably more often than larger ones: for example, 1 appears in about 30% cases, while 9 only about 5% of the time. It turns out that the probability of

digit i is proportional to $\log(\frac{i+1}{i})$. This is called *Benford's law*; apparently the phenomenon was first pointed out by Newcomb, who noticed that the first pages in tables of logarithms, those that contain numbers with leading digit 1, were more worn than the other pages. Benford's law is typically valid for distributions that span several orders of magnitude, and it has been used for detecting frauds in accounting and in science.

3.2. Random Variables and Expectation. The next key notion of probability theory, after a probability space, is a **random variable**. In the above setting, a (real) random variable X on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is simply a P-measurable function $X\colon \Omega \to \mathbb{R}$ (random variables are typically denoted by capital letters). Explicitly, this means that $\{\omega \in \Omega : X(\omega) \le a\}$ is an event for every $a \in \mathbb{R}$ (or equivalently, $X^{-1}(B) \in \mathcal{F}$ for every Borel set $B \subseteq \mathbb{R}$).

Let us consider the probability space $(\Omega_{\mathrm{Sq}}, \mathcal{F}_{\mathrm{Sq}}, \mathrm{P}_{\mathrm{Sq}})$. Examples of random variables on it are the distance of a random point ω from the center of the square, or the area of the triangle spanned by ω and the two bottom corners. For the choice of three independent random points in the square as in Exercise 3.1, one may consider the area of the triangle spanned by the three points as a random variable.



The **expectation** $\mathbb{E}[X]$ of a random variable X is defined as the Lebesgue integral

$$\mathbb{E}[X] := \int_\Omega X(\omega)\,\mathrm{dP}(\omega).$$

Exercise 3.2. Compute the expectation of the first two random variables mentioned above. ✠

Alternative axiomatization: the algebra of random variables and free probability. Having talked about axioms of probability,

we should also mention a different approach, with very interesting generalizations, developed in the last few decades.

Its starting point is that for most problems in probability and statistics, there is a considerable freedom in choosing the underlying probability space, and the specific choice does not really matter—usually it suffices to know the expectations of various algebraic expressions in the random variables involved.

The alternative axiomatization of probability has random variables as elementary objects. They form an algebraic structure (called a *complex commutative $*$-algebra*), obeying suitable axioms, and there is a (linear) operator assigning to every random variable its expectation. When needed, an underlying probability space can then be constructed using suitable representation theorems from mathematical analysis.

This axiomatization has a generalization in which the random variables need not be commutative, in the sense that the expectation of XY may be different from the expectation of YX. In classical probability theory, such noncommutativity appears for matrix-valued random variables. This generalization leads to areas such as the theory of random matrices, quantum probability, or *free probability*, which is an active research field with connections to several other fields (including combinatorics), and which will almost surely be also useful for computer science. We refer, e.g., to [**Tao12**] for an introduction.

4. Literature

Measure and integration are basic and classical areas with plenty of textbooks. For example, Royden [**Roy88**] gives a friendly account, and Tao [**Tao11**] is a modern treatment with many interesting side views. Mattila [**Mat95**] nicely covers geometric aspects of measure theory, such as the construction and properties of the Hausdorff measure and its relatives. Even more textbooks exist for probability theory; here we mention Grimmett and Stirzaker [**GS01**].

Acknowledgment. We would like to thank Vojtěch Kaluža, Aleš Pultr, and Maria Saumell for reading and valuable comments.

Bibliography

[GS01] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, third edition, 2001.

[Mat95] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.

[Roy88] H. L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[Tao11] T. Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

[Tao12] T. Tao. *Topics in Random Matrix Theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[Wag93] S. Wagon. *The Banach–Tarski Paradox*. Cambridge University Press, Cambridge, 1993. With a foreword by Jan Mycielski, Corrected reprint of the 1985 original.