
Chapter 2

Lebesgue Measure

This chapter develops the basic notions of measure theory. They are what is needed to introduce the concepts of measure-preserving transformations, recurrence and ergodicity in Chapter 3. We first develop the theory of Lebesgue measure on the real line. As we shall see, all the basic ideas of measure are already present in the construction of Lebesgue measure on the line. We end with a section on the changes that are necessary to extend our construction of Lebesgue measure from the real line to d -dimensional Euclidean space. The reader is referred to the appendices for mathematical notation not defined here and for basic properties of the real numbers that we use.

2.1. Lebesgue Outer Measure

Lebesgue measure in \mathbb{R} generalizes the notion of length. We will see that the notion of length is generalized to a large class of subsets of the line.

The simplest sets for which we have a good notion of length are the intervals, and they form the starting point for our development of Lebesgue measure. The length of an interval I is denoted by $|I|$. In Theorem 2.1.3 we show that, as expected, the Lebesgue measure of an interval is indeed equal to its length.

Sets will be “measured” by approximating them by countable unions of intervals. We will approximate our sets “from above,” i.e., we will consider unions of intervals *containing* our sets. The idea is to consider all possible countable collections of intervals covering a given set A and to take the sum of the lengths of the respective intervals. Then to obtain the “measure” we take the infimum of all these numbers. Formally, we define the **Lebesgue outer measure** or simply the **outer measure** of a set A in \mathbb{R} by

$$(2.1) \quad \lambda^*(A) = \inf \left\{ \sum_{j=1}^{\infty} |I_j| : A \subset \bigcup_{j=1}^{\infty} I_j, \text{ where } I_j \text{ are bounded intervals} \right\}.$$

The set over which the infimum is taken in (2.1) (i.e., the set of sums of lengths of intervals) is bounded below by 0. Thus, if one such sum in (2.1) is finite, the completeness property of the real numbers implies that $\lambda^*(A)$ is a (finite) nonnegative real number. It may happen that for all intervals covering A , the sum of their lengths is ∞ ; in this case we write $\lambda^*(A) = \infty$ (we will see that this happens if, for example, $A = \mathbb{R}$). With the understanding that the outer measure may be infinite in some cases, we see that the notion of outer measure is defined for every subset A of \mathbb{R} .

It is reasonable to ask what happens if one only takes finite sums in (2.1) instead of infinite sums. This notion is called Jordan content or Peano-Jordan content and it does not yield a countably additive measure.

We are now ready to state some basic properties of the outer measure of a set.

Proposition 2.1.1. *Lebesgue outer measure satisfies the following properties.*

- (1) *The intervals I_j in the definition of outer measure may all be assumed to be open.*
- (2) *For any constant $\delta > 0$, the intervals I_j in the definition of outer measure may all be assumed to be of length less than δ .*
- (3) *For any sets A and B , if $A \subset B$, then $\lambda^*(A) \leq \lambda^*(B)$.*

(4) (*Countable Subadditivity*) For any sequence of sets $\{A_j\}$ in \mathbb{R} it is the case that

$$\lambda^*\left(\bigcup_{j=1}^{\infty} A_j\right) \leq \sum_{j=1}^{\infty} \lambda^*(A_j).$$

Proof. For part (1) let $\alpha(A)$ denote the outer measure of A when computed using only open bounded intervals in the coverings. Clearly, $\lambda^*(A) \leq \alpha(A)$. Now let $\varepsilon > 0$. For any covering $\{I_j\}$ of A let K_j be an open interval containing I_j such that $|K_j| < |I_j| + \frac{\varepsilon}{2^j}, j \geq 1$. Then

$$\sum_{j=1}^{\infty} |K_j| < \sum_{j=1}^{\infty} |I_j| + \sum_{j=1}^{\infty} \frac{\varepsilon}{2^j} = \sum_{j=1}^{\infty} |I_j| + \varepsilon.$$

Taking the infimum of each side gives $\alpha(A) \leq \lambda^*(A) + \varepsilon$, and as this holds for all ε , $\alpha(A) \leq \lambda^*(A)$.

Part (2) follows from the fact that intervals can be subdivided into subintervals without changing the sum of their lengths.

Part (3) follows directly from the definition.

For (4), first note that if for some $j \geq 1$, $\lambda^*(A_j) = \infty$, then there is nothing to prove. Suppose now that for all $j \geq 1$, $\lambda^*(A_j) < \infty$. Let $\varepsilon > 0$. Applying Lemma A.1.1, for each $j \geq 1$ there exist intervals $\{I_{j,k}\}_{k \geq 1}$ such that

$$A_j \subset \bigcup_{k=1}^{\infty} I_{j,k} \quad \text{and} \quad \sum_{k=1}^{\infty} |I_{j,k}| < \lambda^*(A_j) + \frac{\varepsilon}{2^j}.$$

Then

$$A \subset \bigcup_{j=1}^{\infty} \bigcup_{k=1}^{\infty} I_{j,k}$$

and

$$\begin{aligned} \lambda^*(A) &\leq \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |I_{j,k}| < \sum_{j=1}^{\infty} \left(\lambda^*(A_j) + \frac{\varepsilon}{2^j} \right) \\ &= \left(\sum_{j=1}^{\infty} \lambda^*(A_j) \right) + \varepsilon. \end{aligned}$$

Since this holds for all $\varepsilon > 0$, $\lambda^*(A) \leq \sum_{j=1}^{\infty} \lambda^*(A_j)$. \square

We require the following basic lemma about the notion of length.

Lemma 2.1.2. *Let $[a, b]$ be a closed interval and let $\{I_j\}_{j=1}^K$ be any finite collection of open intervals such that $[a, b] \subset \bigcup_{j=1}^K I_j$. Then*

$$|[a, b]| \leq \sum_{j=1}^K |I_j|.$$

Proof. If any I_j is infinite, then $\sum_{j=1}^K |I_j| = \infty$, so we need consider only the case when all the intervals I_j are finite. Write $I_j = (a_j, b_j)$, for $1 \leq j \leq K$. Let j_1 be the smallest integer such that $a \in (a_{j_1}, b_{j_1})$. Let j_2 be the smallest integer such that the previous point $b_{j_1} \in (a_{j_2}, b_{j_2})$. This generates a sequence j_1, j_2, \dots which must terminate, as the collection of intervals from which it comes is finite. From the construction it must terminate at some integer j_ℓ such that $a_{j_\ell} < b < b_{j_\ell}$. Then

$$b - a < b_{j_\ell} - a_{j_1} \leq \sum_{i=1}^{\ell} |I_{j_i}| \leq \sum_{j=1}^K |I_j|.$$

□

The following theorem shows that indeed Lebesgue outer measure generalizes our notion of length.

Theorem 2.1.3. *If I is a (bounded or unbounded) interval, then*

$$\lambda^*(I) = |I|.$$

Proof. Let I be a bounded interval. As I covers itself, we have that $\lambda^*(I) \leq |I|$, so it suffices to show that $|I| \leq \lambda^*(I)$. First assume that $I = [a, b]$, a closed bounded interval. Calculate outer measure using open bounded intervals. Let $\{I_j\}_{j=1}^{\infty}$ be a sequence of open bounded intervals covering I . By Theorem B.1.5, there exists a finite subcollection $\{I_{j_i}\}_{i=1}^K$ of these intervals that covers I . By Lemma 2.1.2,

$$b - a \leq \sum_{i=1}^K |I_{j_i}| \leq \sum_{j=1}^{\infty} |I_j|.$$

This means that $|I| \leq \lambda^*(I)$.

Next consider any bounded interval I . For each $\varepsilon > 0$ choose a closed interval $J_\varepsilon \subset I$ such that $|J_\varepsilon| > |I| - \varepsilon$. Then

$$|I| < |J_\varepsilon| + \varepsilon = \lambda^*(J_\varepsilon) + \varepsilon \leq \lambda^*(I) + \varepsilon.$$

Since this holds for all $\varepsilon > 0$, $|I| \leq \lambda^*(I)$, which is the desired inequality.

Finally, consider an unbounded interval I . Then for any integer $k > 0$ there is a bounded interval $J \subset I$ with $|J| = k$. Therefore, $\lambda^*(I) = \infty$. \square

In closing, define a set N to be a **null set** if its outer measure is zero, i.e., $\lambda^*(N) = 0$. Proposition 2.1.1(4) then implies that countable sets are null sets. An interesting consequence of this and Theorem 2.1.3 is another proof that intervals are not countable. In Section 2.2 we see that the Cantor set provides an example of an uncountable set that is a null set. Cantor sets of positive measure (see Exercise 2.4.5) provide examples of sets that contain no intervals but are not null.

Exercises

- (1) Show that there is no greater generality in the definition of outer measure if the intervals are not restricted to being bounded.
- (2) a) Show that the intervals in the definition of outer measure may be assumed to be closed. b) Show that for any δ , the intervals in the definition of outer measure may be assumed to be open and of length less than δ . c) Show that if A is contained in an interval K , then we can assume that all intervals I_j in the cover are contained in K .
- (3) A **dyadic interval** or 2-adic interval is an interval of the form

$$[k/2^j, (k+1)/2^j)$$

for some integers k and $j \geq 0$. (It is convenient to take them left-closed and right-open.) Show that the intervals I_j in the definition of outer measure may all be assumed to be dyadic intervals.

- (4) For any set $A \subset \mathbb{R}$ and number t , define $A+t = \{a+t : a \in A\}$, the translation of A by t . Show that $\lambda^*(A+t) = \lambda^*(A)$.
- (5) Show that if N is a null set, then for any set A , $\lambda^*(A \cup N) = \lambda^*(A)$.
- (6) Show that if a set is bounded, then its outer measure is finite. Is the converse true?
- (7) Show that the union of countably many null sets is a null set.
- (8) For any $t \in \mathbb{R}$ define $tA = \{ta : a \in A\}$. Show that $\lambda^*(tA) = |t|\lambda^*(A)$.
- (9) Generalize Exercise 3 to the case of q -adic intervals, $q > 2$ (similar to a 2-adic interval but with 2 replaced by q).
- * (10) In the definition of outer measure of a set A , replace countable intervals covering A by finite intervals covering A and call it *outer content*. Determine which of the properties that we have shown for outer measure still hold for outer content.

2.2. The Cantor Set and Null Sets

This section introduces the **Cantor middle-thirds set** K and its basic properties. The Cantor set is a remarkable set that plays a crucial role as a source of examples and counterexamples in analysis and dynamics.

The set will be defined inductively. Let

$$F = [0, 1] \text{ and } G_0 = \left(\frac{1}{3}, \frac{2}{3}\right).$$

We say that G_0 is the *open middle-third* of F . Then set

$$F_1 = F \setminus G_0.$$

F_1 consists of 2^1 closed intervals in F , each of length $\frac{1}{3}$ and denoted by $F[0]$ and $F[1]$. Let $G_1 = \left(\frac{1}{9}, \frac{2}{9}\right) \cup \left(\frac{7}{9}, \frac{8}{9}\right)$, the union of the middle-thirds of each of the subintervals of F_1 . Then set

$$F_2 = F_1 \setminus G_1.$$

F_2 consists of 2^2 closed subintervals of F_1 , each of length $\frac{1}{3^2}$ and denoted, in order from left to right, by $F[00], F[01], F[10], F[11]$. Figure 2.1 shows the first few steps in the construction.

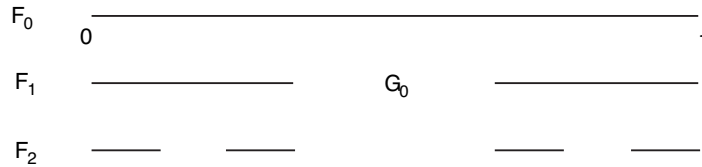


Figure 2.1. First steps in the construction of the Cantor set

Now suppose that F_{n-1} has been defined and consists of 2^{n-1} closed subintervals, each of length $\frac{1}{3^{n-1}}$ and denoted by (ordered from left to right)

$$F[0 \cdots 0], F[0 \cdots 1], \dots, F[1 \cdots 1].$$

Let G_{n-1} be the union of the open middle-thirds (each of length $\frac{1}{3^n}$) of each of the 2^{n-1} closed subintervals of F_{n-1} . Then set

$$F_n = F_{n-1} \setminus G_{n-1},$$

a union of 2^n closed subintervals, each of length $\frac{1}{3^n}$. The Cantor set K is defined by

$$K = \bigcap_{n=1}^{\infty} F_n.$$

It follows from this definition that all endpoints of the subintervals in F_n belong to K . It might seem that “most” points of $[0, 1]$ have been removed, but in fact, as we shall see in the theorem below, there are uncountably many points that are left in K . The next exercise provides an alternative way to think about K .

Question. Show that the Cantor set is also given by

$$K = F \setminus \left(\bigcup_{n=0}^{\infty} G_n \right).$$

Thus, the Cantor set is the set of points in $[0, 1]$ that is obtained after removing all the open intervals comprising the sets G_n .

It is interesting to note the following characterization of a null set; its proof is left to the reader.

Lemma 2.2.1. *A set N is a null set if and only if for any $\varepsilon > 0$ there exists a sequence of intervals I_j such that*

$$N \subset \bigcup_{j=1}^{\infty} I_j \text{ and } \sum_{j=1}^{\infty} |I_j| < \varepsilon,$$

where $|I_j|$ denotes the length of the interval I_j .

The idea is that the set N can be covered by intervals such that the sum of their lengths can be made arbitrarily small. A set consisting of a single point p is clearly a null set, as it is covered by $(p - \varepsilon, p + \varepsilon)$ for all $\varepsilon > 0$. (As intervals are allowed to be points, a simpler proof could be obtained by taking the interval covering $\{p\}$ to be $I = [p, p]$, but we have given a proof that also works with the more restrictive definition requiring the covering intervals to be open, or of positive length. Exercise 3 shows that both definitions are equivalent.) The reader may verify that countable sets are null sets. Surprisingly, there exist uncountable sets that are null sets. An important example of this is provided by the Cantor set.

Theorem 2.2.2. *The Cantor set K is a null set and an uncountable closed subset of $[0, 1]$. Furthermore, K contains no positive length intervals and is perfect, i.e., every point of K is an accumulation point of K .*

Proof. We know that

$$K = \bigcap_{n=1}^{\infty} F_n,$$

where F_n is a union of 2^n closed intervals, each of length $1/3^n$. Given $\varepsilon > 0$ choose n so that $(2/3)^n < \varepsilon$. Then the sum of the lengths of the intervals comprising F_n is less than ε . Therefore K is covered by a finite union of intervals whose total length is less than ε , showing that K is a null set. Also, K is closed as it is an intersection of closed sets.

If I is an interval contained in K , then $I \subset F_n$ for all $n > 0$. As I is an interval, it must be contained in one of the subintervals of F_n , each of length $1/3^n$. So $|I| < 1/3^n$, for all $n > 0$. Therefore K contains no intervals of positive length.

To show that K is uncountable we define the following function. First, represent $x \in [0, 1]$ in binary form as

$$x = \sum_{i=1}^{\infty} \frac{x_i}{2^i}, \text{ where } x_i \in \{0, 1\}.$$

For example $\frac{1}{3}$ is represented by

$$\frac{0}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \cdots$$

since

$$\sum_{i=1}^{\infty} \frac{1}{2^{2i}} = \sum_{i=1}^{\infty} \frac{1}{4^i} = \frac{1}{3}.$$

This representation is unique if we assume that it does not end in an infinite sequence of 1's. More precisely, let D consist of all the numbers in $[0, 1]$ of the form $\frac{k}{2^n}$, for integers $k \geq 0, n > 0$, and let $I_0 = [0, 1] \setminus D$. D is countable and consists precisely of the numbers in $[0, 1]$ that have more than one representation in binary form. (For example $\frac{1}{2}$ may be written in one way with $x_1 = 1$ and $x_i = 0$ for $i \geq 2$ and in another way with $x_1 = 0$ and $x_i = 1$ for $i \geq 2$.) So each $x \in X_0$ has a unique binary representation. For

$$x = \sum_{i=1}^{\infty} \frac{x_i}{2^i} \in X_0$$

define the map $\phi : X_0 \rightarrow K$ by

$$(2.2) \quad \phi(x) \in \bigcap_{n=1}^{\infty} F[x_1 x_2 \cdots x_n].$$

To show that ϕ is a function we need to verify that the intersection in (2.2) consists of a single point. Note that since the sets $F[x_1 x_2 \cdots x_n]$ are closed and are contained in F_n and since

$$\lim_{n \rightarrow \infty} \lambda(F_n) = 0,$$

the intersection $\bigcap_{n=1}^{\infty} F[x_1 x_2 \cdots x_n]$ contains a unique point, and this point must be in K . So ϕ is a well-defined function from X_0 into K . To show that K is uncountable it suffices to show that ϕ is one-to-one. (ϕ is also onto but that is not needed here and is left to the reader as an exercise.) Let $x, y \in X_0$. Then $x = \sum_{i=1}^{\infty} x_i/2^i$ and $y = \sum_{i=1}^{\infty} y_i/2^i$. Suppose $x \neq y$. Then there exists some $k > 0$ such that $x_k \neq y_k$.

This means that the sets $F[x_1x_2 \cdots x_k]$ and $F[y_1y_2 \cdots y_k]$ are disjoint, which implies that $\phi(x) \neq \phi(y)$, so ϕ is one-to-one. It follows that the cardinality of K must be at least that of X_0 , so K must be uncountable.

To see that K is perfect, for any $x \in K$ and for any $n > 0$ let a_n be any endpoint, different from x , of the subinterval in F_n that contains x . Then $\{a_n\}$ is a sequence of points of K different from x and converging to x (as $|x - a_n| < 1/2^n$). So K is perfect. \square

A set is said to be **nowhere dense** if its closure has empty interior, i.e., its closure contains no open sets. This is the same as saying that there is no (nonempty) interval in which the set is dense. Evidently, any finite set in \mathbb{R} is nowhere dense, but infinite sets may also be nowhere dense (such as \mathbb{Z} or $\{1/n\}_{n>0}$). A set is said to be **totally disconnected** if its *connected components* (not defined here) are just points. It can be shown for the case of subsets of \mathbb{R} that a set is totally disconnected if and only if it contains no positive length intervals, and we adopt this as our definition of totally disconnected for a subset of \mathbb{R} . Evidently, a closed set in \mathbb{R} is nowhere dense if and only if it is totally disconnected. We have just shown that the Cantor middle-thirds set is a closed, bounded, perfect and totally disconnected subset of \mathbb{R} . Any subset of \mathbb{R} satisfying these properties is called a **Cantor set**. The notion of a Cantor set can be defined for more general sets, but one needs the notion of a *homeomorphism* between topological spaces. Then using these notions (not defined in this book) a more general Cantor set can be defined as any topological space that is homeomorphic to the Cantor middle-thirds set.

We have observed that the notion of Lebesgue measure zero coincides with the notion of null set. Using the notion of measure, another way to see that K is a set of measure zero is to show that its complement G in $[0, 1]$ has measure 1. This is a simple computation as the sum of the lengths of the intervals comprising G is

$$\lambda(G) = \sum_{n=0}^{\infty} \frac{2^n}{3^{n+1}} = 1.$$

We shall see in Theorem 2.4.1 that if $K \sqcup G = [0, 1]$, then the measure of K is $1 - \lambda(G) = 0$. One can intuitively think that K has measure

zero as it is the set that remains in $[0, 1]$ after removing a set G of measure 1. This is also the starting point for constructing other Cantor sets that are not null; they are obtained after removing a disjoint countable collection of intervals such that the sum of their lengths add to less than 1 (see Exercise 2.4.(5)).

Define a transformation $T : \mathbb{R} \rightarrow \mathbb{R}$ by

$$T(x) = \begin{cases} 3x & \text{if } x \leq \frac{1}{2}; \\ 3 - 3x & \text{if } x > \frac{1}{2}. \end{cases}$$

A set that is interesting to consider in dynamics is the set of points x whose positive orbit under T , i.e., the set $\{T^n(x) : n \geq 0\}$, is bounded. So define Λ to be such that if $x \in \Lambda$, then the positive orbit $\{T^n(x)\}_{n \geq 0}$ is bounded. Note that if $x \in \Lambda$, then $T(x) \in \Lambda$. For the case of the transformation T first observe that if $x < 0$, then $T(x) = 3x < 0$, and by induction $T^n(x) = 3^n x$ for all $n > 0$. It follows that if $x < 0$, or $T^k(x) < 0$ for some k , then $T^n(x) \rightarrow -\infty$ as $n \rightarrow \infty$. Next observe that if $T^n(x) > 1$ for some $n \geq 0$, then $T^{n+1}(x) = T(T^n(x)) < 0$, so $x \notin \Lambda$. Clearly, $0 \in \Lambda$. Thus Λ is characterized by

$$(2.3) \quad \Lambda = \{x \in [0, 1] : T^n(x) \in [0, 1] \text{ for all } n \geq 0\}.$$

Using (2.3), we now identify Λ with the middle-thirds Cantor set K . For this write each point $x \in K$ in its ternary representation. If $x \in K \cap [0, \frac{1}{2}]$, then

$$x = \sum_{i=2}^{\infty} \frac{x_i}{3^i}.$$

So,

$$T(x) = 3x = \sum_{i=2}^{\infty} \frac{x_i}{3^{i-1}} = \sum_{i=1}^{\infty} \frac{x_{i+1}}{3^i}.$$

This shows that $T(x)$ is in K . In addition it describes the effect of T on the point x : we see that if the digits in the ternary representation of x are $x_1 x_2 \cdots$, then the digits in the ternary representation of $T(x)$ are $x_2 x_3 \cdots$. In other words, T shifts the representation of x to the

left. Now if $x \in K \cap [\frac{1}{2}, 1]$, then

$$x = \frac{2}{3} + \sum_{i=2}^{\infty} \frac{x_i}{3^i}.$$

So,

$$T(x) = 3 - 3\left(\frac{2}{3} + \sum_{i=2}^{\infty} \frac{x_i}{3^i}\right) = \sum_{i=1}^{\infty} \frac{2 - x_{i+1}}{3^i}.$$

So, $T(x) \in K$. In a similar way one can show that if $x \in [0, 1] \setminus K$, then $T^k(x) > 1$ for some $k > 0$, so $x \notin \Lambda$. One concludes that $K = \Lambda$.

Exercises

- (1) Prove Lemma 2.2.1.
- (2) Let F_n be as in the construction of the Cantor set K . Show that for all n , the endpoints of the closed subintervals in F_n belong to the Cantor set.
- (3) Show that in Lemma 2.2.1 the sets I_j may be assumed to be nonempty open intervals.
- (4) Modify the construction of the Cantor middle-thirds set in the following way. At the first stage remove a central interval of length $\frac{1}{5}$ and at the n^{th} stage, instead of removing the open middle-thirds, remove the open middle-fifth of each subinterval. Show that in this way you obtain a Cantor set that is a null set.
- (5) Recall that every $x \in [0, 1)$ can be written in ternary expansion as $x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}$, where $a_i \in \{0, 1, 2\}$. Show that the Cantor set K is precisely $K = \{\sum_{i=1}^{\infty} \frac{a_i}{3^i} \text{ where } a_i \in \{0, 2\}\}$.
- (6) Show that if K is the middle-thirds Cantor set, then $K + K = [0, 2]$, where $K + K = \{z : z = x + y \text{ for some } x, y \in K\}$.
- (7) Show that the function ϕ in the proof of Theorem 2.2.2 is onto.
- * (8) A real number x is said to be a *Liouville number* if it is irrational and for any integer $n > 0$ there exist integers p and $q > 1$ such that

$$\left|x - \frac{p}{q}\right| < \frac{1}{q^n}.$$

Show that the set of Liouville numbers is a null set.

- * (9) (The Cantor function) Construct a function $\psi : [0, 1] \rightarrow [0, 1]$ that is continuous, monotone increasing (i.e., if $x \leq y$, then $f(x) \leq f(y)$) in $[0, 1]$ and such that ψ is constant on each interval in the complement of K and $\psi(K) = [0, 1]$. This is called the Cantor ternary function. (Hint: Represent x in ternary form and $\psi(x)$ in binary form. The value of ψ on the intervals G_n should be constant.)

2.3. Lebesgue Measurable Sets

A set A in \mathbb{R} is said to be **Lebesgue measurable** or **measurable** if for any $\varepsilon > 0$ there is an open set $G = G_\varepsilon$ such that

$$A \subset G \text{ and } \lambda^*(G \setminus A) < \varepsilon.$$

Informally, we see from the definition that measurable sets are those that are “well-approximated” from above by open sets. We first obtain some examples of measurable sets.

Proposition 2.3.1. *Open sets and null sets are measurable.*

Proof. If A is open, then for any $\varepsilon > 0$, G can be taken to be A , so open sets are clearly measurable. Now let N be a null set, i.e., suppose $\lambda^*(N) = 0$. Then for any $\varepsilon > 0$ there exists a sequence of open intervals $\{I_j\}$ whose union covers N and such that $\sum_{j=1}^{\infty} |I_j| < \varepsilon$. Then $G = \bigcup_{j=1}^{\infty} I_j$ is open and satisfies

$$\lambda^*(G \setminus N) \leq \lambda^*(G) \leq \sum_{j=1}^{\infty} |I_j| < \varepsilon.$$

Hence N is measurable. □

Our goal in this section is to show that countable unions, countable intersections and complements of measurable sets are measurable. Together with the fact that open sets and null sets are measurable, this yields a large class of subsets of the line that are measurable; the sets in this class will be defined later as the Lebesgue measurable sets.

Showing that the countable union of measurable sets is measurable is a rather straightforward consequence of the definition, as the following proof of Proposition 2.3.2 shows. The proof for countable intersections is more difficult, and for us it will follow from Proposition 2.3.2 and the fact that complements of measurable sets are measurable. The crucial part for this last fact is to show that closed sets are measurable. In the process, we also obtain a useful characterization of measurable sets (Lemma 2.3.7).

Proposition 2.3.2. *The countable union of measurable sets is measurable.*

Proof. Let $\{A_n\}_{n \geq 1}$ be a sequence of measurable sets and write $A = \bigcup_{n=1}^{\infty} A_n$. Let $\varepsilon > 0$. For each n there exists an open set G_n such that $A_n \subset G_n$ and

$$\lambda^*(G_n \setminus A_n) < \frac{\varepsilon}{2^n}.$$

Let $G = \bigcup_{n=1}^{\infty} G_n$; then G is open and covers A . Since

$$G \setminus A \subset \bigcup_{n=1}^{\infty} (G_n \setminus A_n),$$

then, using countable subadditivity,

$$\lambda^*(G \setminus A) \leq \lambda^*\left(\bigcup_{n=1}^{\infty} (G_n \setminus A_n)\right) \leq \sum_{n=1}^{\infty} \lambda^*(G_n \setminus A_n) < \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon.$$

Therefore A is measurable. \square

The following lemma is an important fact that is true in greater generality (see Theorem 2.4.1), but we need this special case here to prove that closed sets are measurable.

Lemma 2.3.3. *Let $\{E_j\}_{j=1}^N$ be a finite collection of disjoint closed bounded sets. Then*

$$\lambda^*\left(\bigsqcup_{j=1}^N E_j\right) = \sum_{j=1}^N \lambda^*(E_j).$$

Proof. By countable subadditivity we only need to show that

$$\lambda^*\left(\bigsqcup_{j=1}^N E_j\right) \geq \sum_{j=1}^N \lambda^*(E_j).$$

First assume $N = 2$; the general case follows immediately by induction. By Lemma B.1.6 there exists a number $\delta > 0$ so that every interval of length less than δ cannot have a nonempty intersection with both E_1 and E_2 . As $E_1 \cup E_2$ is bounded, $\lambda^*(E_1 \cup E_2) < \infty$. Let $\varepsilon > 0$. Use Proposition 2.1.1 to find a sequence of intervals $\{I_j\}_{j \geq 1}$, with $|I_j| < \delta$, whose union covers $E_1 \cup E_2$ and such that

$$\sum_{j=1}^{\infty} |I_j| < \lambda^*(E_1 \cup E_2) + \varepsilon.$$

Let $\Gamma = \{j \geq 1 : I_j \cap E_1 \neq \emptyset\}$. Then

$$E_1 \subset \bigcup_{j \in \Gamma} I_j \text{ and } E_2 \subset \bigcup_{j \in \Gamma^c} I_j.$$

Therefore,

$$\lambda^*(E_1) + \lambda^*(E_2) \leq \sum_{j \in \Gamma} |I_j| + \sum_{j \in \Gamma^c} |I_j| = \sum_{j=1}^{\infty} |I_j| < \lambda^*(E_1 \cup E_2) + \varepsilon.$$

As this holds for all $\varepsilon > 0$, then $\lambda^*(E_1) + \lambda^*(E_2) \leq \lambda^*(E_1 \cup E_2)$ and this completes the proof. \square

We now prove a technical lemma to be used in Lemma 2.3.5 but also of interest in its own right (a generalization of this lemma is offered in Exercise 2.4.2). The new idea in Lemma 2.3.4 is that the intervals are not necessarily disjoint but are allowed to intersect at their endpoints. While the lemma is stated for any bounded intervals, we will only apply it in the case of bounded closed intervals.

Lemma 2.3.4. *Let $\{I_j\}_{j=1}^N$ be a finite collection of bounded intervals that are nonoverlapping. Then*

$$\lambda^*\left(\bigcup_{j=1}^N I_j\right) = \sum_{j=1}^N \lambda^*(I_j).$$

Proof. Since the intervals $\{I_j\}$ are nonoverlapping, for each $\varepsilon > 0$ there exists a closed interval $I'_j \subset I_j$ such that $\lambda^*(I'_j) \geq \lambda^*(I_j) - \frac{\varepsilon}{N}$,

with $\{I'_j\}$ disjoint. Using Lemma 2.3.3 we obtain that

$$\begin{aligned} \lambda^*\left(\bigcup_{j=1}^N I_j\right) &\geq \lambda^*\left(\bigsqcup_{j=1}^N I'_j\right) = \sum_{j=1}^N \lambda^*(I'_j) \\ &\geq \sum_{j=1}^N \lambda^*(I_j) - \varepsilon. \end{aligned}$$

Letting $\varepsilon \rightarrow 0$, $\lambda^*(\bigcup_{j=1}^N I_j) \geq \sum_{j=1}^N \lambda^*(I_j)$. The reverse inequality follows from subadditivity. \square

Lemma 2.3.5 is true in greater generality (see Corollary 2.4.2) but we need it now to show that closed sets are measurable.

Lemma 2.3.5. *If F is a bounded closed set and G is an open set such that $F \subset G$, then*

$$\lambda^*(G \setminus F) = \lambda^*(G) - \lambda^*(F).$$

Proof. Since $G = (G \setminus F) \cup F$ and $\lambda^*(F) < \infty$, by countable subadditivity $\lambda^*(G \setminus F) \geq \lambda^*(G) - \lambda^*(F)$. To prove the other inequality we observe that $G \setminus F$ is open, so there exists a sequence of nonoverlapping, closed bounded intervals $\{I_j\}$ such that $G \setminus F = \bigcup_{j=1}^{\infty} I_j$. Thus, for all $N > 1$,

$$G \supset \left(\bigcup_{j=1}^N I_j\right) \sqcup F.$$

As both $\bigcup_{j=1}^N I_j$ and F are disjoint, bounded and closed, applying Lemma 2.3.3 and then Lemma 2.3.4, we obtain that

$$\begin{aligned} \lambda^*(G) &\geq \lambda^*\left(\bigcup_{j=1}^N I_j\right) + \lambda^*(F) \\ &= \sum_{j=1}^N \lambda^*(I_j) + \lambda^*(F). \end{aligned}$$

Since this holds for all N ,

$$\lambda^*(G) \geq \sum_{j=1}^{\infty} \lambda^*(I_j) + \lambda^*(F) \geq \lambda^*(G \setminus F) + \lambda^*(F).$$

\square

We are now ready to prove the following proposition. This, together with Lemma 2.3.7, is used later to prove that the complement of a measurable set is measurable.

Proposition 2.3.6. *Closed sets are measurable.*

Proof. Let F be a closed set. First assume that F is bounded, hence of finite outer measure. Thus, for $\varepsilon > 0$ there exist open intervals I_j such that $F \subset \bigcup_{j=1}^{\infty} I_j$ and $\lambda^*(\bigcup_{j=1}^{\infty} I_j) < \lambda^*(F) + \varepsilon$. Let $G = \bigcup_{j=1}^{\infty} I_j$. Then G is open and contains F and by Lemma 2.3.5 $\lambda^*(G \setminus F) = \lambda^*(G) - \lambda^*(F) < \varepsilon$, so F is measurable.

In the general case, write $F_n = F \cap [-n, n]$. Then F_n is closed and bounded and therefore measurable. By Proposition 2.3.2, the union $\bigcup_{n=-\infty}^{\infty} F_n = F$ is measurable. \square

The following lemma tells us that a measurable set is “almost” a countable intersection of open sets; its converse is also true and is a consequence of Theorem 2.3.8. A set that is a countable intersection of open sets, such as the set G^* in Lemma 2.3.7, is called a \mathcal{G}_δ set. (In this notation \mathcal{G} stands for open and δ for intersection.)

Lemma 2.3.7. *If a set A is measurable, then there exists a \mathcal{G}_δ set G^* and a null set N such that*

$$A = G^* \setminus N.$$

Proof. Let A be measurable. For each $\varepsilon_n = \frac{1}{n} > 0$ there exists an open set G_n with $A \subset G_n$ and $\lambda^*(G_n \setminus A) < \varepsilon_n$. Let $G^* = \bigcap_n G_n$. Then $A \subset G^*$, and for all n ,

$$\lambda^*(G^* \setminus A) \leq \lambda^*(G_n \setminus A) < \frac{1}{n}.$$

Thus $\lambda^*(G^* \setminus A) = 0$. If $N = G^* \setminus A$, then $A = G^* \setminus N$ and this completes the proof. \square

We now state the main result of this section. Parts (1) and (2) have already been shown but are mentioned again for completeness.

Theorem 2.3.8. *The collection of measurable sets satisfies the following properties.*

- (1) *The empty set and the set of reals \mathbb{R} are measurable.*

- (2) *A countable union of measurable sets is measurable.*
 (3) *The complement of a measurable set is measurable.*
 (4) *A countable intersection of measurable sets is measurable.*

Proof. Parts (1) and (2) have already been shown in Proposition 2.3.1 and Proposition 2.3.2. For part (3), let A be a measurable set. We know that

$$A = G^* \setminus N,$$

where G^* is a \mathcal{G}_δ set and N is a null set. Write $G^* = \bigcap_{n=1}^{\infty} G_n$, where the sets G_n are open. Then

$$A^c = (G^* \cap N^c)^c = \left(\bigcap_{n=1}^{\infty} G_n \right)^c \cup N = \bigcup_{n=1}^{\infty} G_n^c \cup N.$$

Since the sets G_n^c are closed, and N is measurable, then it follows that A^c is measurable.

Part (4) follows from De Morgan's laws (see Exercise A.6). Indeed, if A_n is a sequence of measurable sets, $(\bigcap_{n=1}^{\infty} A_n)^c = \bigcup_{n=1}^{\infty} A_n^c$, which is measurable by part (2). \square

We end with another useful characterization of measurable sets.

Lemma 2.3.9. *A set A is measurable if and only if for any $\varepsilon > 0$ there is a closed set F such that $F \subset A$ and $\lambda^*(A \setminus F) < \varepsilon$.*

Proof. Let A be measurable. By Theorem 2.3.8(3) the set A^c is measurable, so for $\varepsilon > 0$ there exists an open set G such that $A^c \subset G$ and $\lambda^*(G \setminus A^c) < \varepsilon$. One can verify that $G^c \subset A$ and $G \setminus A^c = A \setminus G^c$. Then the set $F = G^c$ is closed and $\lambda^*(A \setminus F) < \varepsilon$. For the converse note that if A satisfies the condition of the lemma, then a similar argument shows that A^c is measurable, so A must be measurable. \square

Exercises

- (1) Show that if A is measurable, then for any t in \mathbb{R} , $A + t = \{a + t : a \in A\}$ and $tA = \{ta : a \in A\}$ are measurable. Conclude that then $\lambda(A + t) = \lambda(A)$ and $\lambda(tA) = |t|\lambda(A)$.
- (2) Show that if A is a null set, then $A^2 = \{a^2 : a \in A\}$ is also a null set.

- (3) Let A be any set. Show that if there is a measurable set B that differs from A by a null set, i.e., $\lambda^*(A \triangle B) = 0$, then A is measurable.
- (4) Show the converse of Lemma 2.3.7, i.e., show that if $A = G^* \setminus N$ where G^* is a \mathcal{G}_δ set and N is a null set, then A is measurable.
- (5) Show that a set A is measurable if and only if there exist a set F^* and a set N such that F^* is a countable union of closed sets, N is a null set and $A = F^* \cup N$. A set that is a countable union of closed sets is called an \mathcal{F}_σ set. (In this notation \mathcal{F} stands for closed and σ for union.)
- (6) Show that every closed set is a \mathcal{G}_δ and every open set is an \mathcal{F}_σ .
- (7) Show that A is measurable if and only if for any $\varepsilon > 0$ there is a closed set F and an open set G such that $F \subset A \subset G$ and $\lambda^*(G \setminus F) < \varepsilon$.
- (8) Let A be a bounded set. Show that A is measurable if and only if for any $\varepsilon > 0$ there is a closed set F such that $F \subset A$ and $\lambda^*(F) > \lambda^*(A) - \varepsilon$. What is the difference between this characterization and Lemma 2.3.9?
- * (9) (*Carathéodory's Criterion*) Show that a set A is a measurable set if and only if for any set B it is the case that $\lambda^*(B) = \lambda^*(B \cap A) + \lambda^*(B \cap A^c)$. (This is studied in the context of arbitrary measure spaces in Section 2.8.)

2.4. Countable Additivity

When restricted to the Lebesgue measurable sets \mathfrak{L} , Lebesgue outer measure λ^* is denoted by λ and called the **Lebesgue measure** on \mathbb{R} ; so $\lambda(A) = \lambda^*(A)$ for all measurable sets A .

We next show one of the most important properties of Lebesgue measure. This property is called countable additivity and is what characterizes a *measure* as defined in the next section.

Theorem 2.4.1 (Countable Additivity). *If $\{A_n\}_{n=1}^{\infty}$ is a sequence of disjoint measurable sets, then*

$$\lambda\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \lambda(A_n).$$

Proof. Let

$$A = \bigsqcup_{n=1}^{\infty} A_n.$$

By countable subadditivity, it suffices to show the following inequality:

$$\sum_{n=1}^{\infty} \lambda(A_n) \leq \lambda(A).$$

We do this first for the case when A is bounded (in fact, we only use that each A_n is bounded). In this case, by Lemma 2.3.9, for each $\varepsilon > 0$ and $n \geq 1$ there exists a closed set $F_n \subset A_n$ such that $\lambda(A_n \setminus F_n) < \frac{\varepsilon}{2^n}$. As $A_n = (A_n \setminus F_n) \cup F_n$ and $\lambda(A_n) < \infty$, we have

$$\lambda(A_n) < \lambda(F_n) + \frac{\varepsilon}{2^n}.$$

So for every integer $N > 1$,

$$\sum_{n=1}^N \lambda(A_n) < \sum_{n=1}^N \lambda(F_n) + \varepsilon.$$

Since the sets F_n are disjoint, closed and bounded, using Lemma 2.3.3,

$$\lambda\left(\bigsqcup_{n=1}^N F_n\right) = \sum_{n=1}^N \lambda(F_n).$$

Therefore,

$$\sum_{n=1}^N \lambda(A_n) < \lambda\left(\bigsqcup_{n=1}^N F_n\right) + \varepsilon \leq \lambda(A) + \varepsilon.$$

Taking limits as $N \rightarrow \infty$ we obtain

$$\sum_{n=1}^{\infty} \lambda(A_n) \leq \lambda(A) + \varepsilon.$$

Letting $\varepsilon \rightarrow 0$ completes the proof of this part.

Finally, if A is not bounded, then for any integer i and any $n \geq 1$ write

$$B_{n,i} = A_n \cap [i, i+1).$$

It follows that for each i , $[i, i+1) \cap A = \bigsqcup_{n=1}^{\infty} B_{n,i}$, a disjoint union. Since $[i, i+1) \cap A$ is bounded, the first part yields

$$\lambda([i, i+1) \cap A) = \sum_{n=1}^{\infty} \lambda(B_{n,i}).$$

Now for any $N > 0$,

$$\lambda(A) \geq \lambda\left(\bigsqcup_{i=-N}^N [i, i+1) \cap A\right) = \sum_{i=-N}^N \lambda([i, i+1) \cap A).$$

Therefore,

$$\begin{aligned} \lambda(A) &\geq \sum_{i=-\infty}^{\infty} \lambda([i, i+1) \cap A) = \sum_{i=-\infty}^{\infty} \sum_{n \geq 1} \lambda(B_{i,n}) \\ &= \sum_{n=1}^{\infty} \sum_{i=-\infty}^{\infty} \lambda(B_{i,n}) = \sum_{n=1}^{\infty} \lambda(A_n), \end{aligned}$$

which completes the proof. \square

The following is a useful corollary whose proof is left to the reader.

Corollary 2.4.2. *Let A and B be measurable sets such that $B \subset A$ and $\lambda(B) < \infty$. Then $\lambda(A \setminus B) = \lambda(A) - \lambda(B)$.*

Exercises

- (1) Prove Corollary 2.4.2. Show that if A and B are measurable sets with $\lambda(B) < \infty$, then $\lambda(A \setminus B) \geq \lambda(A) - \lambda(B)$.
- (2) Define a sequence of measurable sets $\{A_n\}_{n \geq 1}$ to be **almost disjoint** if $\lambda(A_n \cap A_m) = 0$ for all $n \neq m$. Show that if $\{A_n\}_{n \geq 1}$ are almost disjoint, then $\lambda(\bigcup_{n \geq 1} A_n) = \sum_{n \geq 1} \lambda(A_n)$.
- (3) Let A be a set of finite outer measure. Show that the set A is measurable if and only if for any $\varepsilon > 0$ there is a set H such that H is a finite union of bounded intervals and $\lambda^*(A \triangle H) < \varepsilon$.

- (4) Show that any collection of disjoint sets of positive measure is countable. (Note that, as we saw in Section 2.2, positive measure sets need not contain intervals.)
- (5) Modify the construction of the Cantor middle-thirds set in the following way. At the first stage remove a central interval of length $\frac{1}{6}$ and at the n^{th} stage, instead of removing the open middle-thirds, remove slightly smaller intervals so that in the end the total measure of the removed intervals is $\frac{1}{2}$. Show that in this way you obtain a Cantor set that is not a null set. What is the measure of this set?
- (6) Modify the construction of the previous exercise to obtain a set of measure α for any $0 \leq \alpha < 1$ and show that this set is bounded, perfect and totally disconnected.
- (7) (Measure-theoretic union) Let $\{A_\alpha\}_{\alpha \in \Gamma}$ be an arbitrary collection of measurable sets. Show that there exists a measurable set A (called a *measure-theoretic union*) such that $A \subset \bigcup A_\alpha$ and $\lambda(A_\alpha \setminus A) = 0$ for all $\alpha \in \Gamma$. (Hint: First assume that all sets are in $[0, 1]$ and consider the collection of all countable unions of elements from $\{A_\alpha\}_{\alpha \in \Gamma}$.)

2.5. Sigma-Algebras and Measure Spaces

The unit interval with Lebesgue measure is the prototype of a finite measure space. It is the most important measure space that we study. The collection of Lebesgue measurable subsets of the unit interval is one of the important examples of a σ -algebra.

Theorem 2.3.8 states that the collection of Lebesgue measurable sets is a nonempty collection of subsets of \mathbb{R} that is closed under countable unions, countable intersections and complements. A nonempty collection of sets with these properties is called a σ -algebra. This concept plays a crucial role in the development of the general theory of measure. While we give the general definition, our emphasis will be on three kinds of σ -algebras: the Lebesgue measurable sets in \mathbb{R} , the Borel sets in \mathbb{R} , and the collection of all subsets of a finite or countable set X . As we shall see, each of these σ -algebras induces a

σ -algebra on the subsets of a Lebesgue set, a Borel set or a subset of X .

Let X be a nonempty set (usually a measurable subset of \mathbb{R} or \mathbb{R}^d). A σ -**algebra** on X is a collection \mathcal{S} of subsets of X such that

- (1) \mathcal{S} is nonempty;
- (2) \mathcal{S} is closed under complements, i.e., whenever $A \in \mathcal{S}$, then $A^c \in \mathcal{S}$ (here $A^c = X \setminus A$);
- (3) \mathcal{S} is closed under countable unions, i.e., whenever $A_n \in \mathcal{S}, n \geq 1$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$.

Since $\bigcap_{n=1}^{\infty} A_n = (\bigcup_{n=1}^{\infty} A_n^c)^c$, a σ -algebra is closed under countable intersections. As \mathcal{S} must contain at least one element, say A , and $X = A \cup A^c$, it follows that X , and hence \emptyset , are always in \mathcal{S} . We have already seen that the collection of Lebesgue measurable subsets of \mathbb{R} is a σ -algebra. There are two σ -algebras that exist on any set X . The first is $\{\emptyset, X\}$, the smallest σ -algebra of subsets of X , called the **trivial σ -algebra**. The other one is the collection of all subsets of X or the **power set** of X , denoted by $\mathcal{P}(X)$, the largest σ -algebra (i.e., contains any other σ -algebra) of subsets of X , called the **improper σ -algebra**. The improper σ -algebra on a set X will be of interest when X is a finite or countable set.

Question. Show that if \mathcal{S} is a collection of subsets of a set X that contains X and is closed under set-differences (i.e., if $A, B \in \mathcal{S}$, then $A \setminus B \in \mathcal{S}$) and countable unions, then \mathcal{S} is a σ -algebra.

Example. Let $X = \{a, b, c, d\}$. Then the collection \mathcal{S} defined by

$$\mathcal{S} = \{\emptyset, X, \{a\}, \{b, c, d\}, \{a, b\}, \{c, d\}, \{a, c, d\}, \{b\}\}$$

is a σ -algebra on X . Also, if we set $Y = \{a, c, d\}$ and let $\mathcal{S}(Y)$ denote the collection $\{A \cap Y : A \in \mathcal{S}\}$, then

$$\mathcal{S}(Y) = \{\emptyset, Y, \{a\}, \{c, d\}\}$$

is a σ -algebra on Y . Also note that $\mathcal{S}(Y) = \{A : A \subset Y \text{ and } A \in \mathcal{S}\}$. Now, if we let $Z = \{a, b, c\}$, then Z is not in \mathcal{S} and

$$\mathcal{S}(Z) = \{\emptyset, Z, \{a\}, \{b, c\}, \{a, b\}, \{c\}, \{a, c\}, \{b\}\}$$

is also a σ -algebra on Z . But note that in this case the collection $\{A : A \subset Z \text{ and } A \in \mathcal{S}\} = \{\emptyset, \{a\}, \{a, b\}, \{b\}\}$ is not a σ -algebra on Z .

The proof of the following proposition is left to the reader.

Proposition 2.5.1. *Let X be a nonempty set and let \mathcal{S} be a σ -algebra on X .*

- (1) *If $Y \subset X$, then the collection of sets restricted to Y defined by $\mathcal{S}(Y) = \{A \cap Y : A \in \mathcal{S}\}$ is a σ -algebra on Y . ($\mathcal{S}(Y)$ is also denoted by $\mathcal{S} \cap Y$.)*
- (2) *If $Y \in \mathcal{S}$, then*

$$\mathcal{S}(Y) = \{A : A \subset Y \text{ and } A \in \mathcal{S}\}.$$

The collection of all Lebesgue measurable sets in \mathbb{R} is denoted by \mathfrak{L} . If X is a Lebesgue measurable subset of \mathbb{R} , the set of Lebesgue measurable sets contained in X is denoted by $\mathfrak{L}(X)$. By Proposition 2.5.1 $\mathfrak{L}(X)$ is a σ -algebra of subsets of X .

It is important to note that there exist subsets of the reals that are not Lebesgue measurable; a construction of a nonmeasurable set is given in Section 3.2. In fact, it can further be shown that every set of positive Lebesgue outer measure contains a non-Lebesgue measurable subset (see Exercise 3.11.1 or Oxtoby [56, Ch. 5]).

Lebesgue measure is defined on the σ -algebra of measurable sets. As we shall see, sometimes we may find it useful to consider other “measures” such as a multiple of Lebesgue measure. The important property that Lebesgue measure enjoys is that it is countably additive. To make this precise we introduce the following definition. Let X be a nonempty set and \mathcal{S} a σ -algebra on X . A **measure on \mathcal{S}** is a function μ defined on \mathcal{S} and with values in $[0, \infty]$ that satisfies the following two properties:

- (1) $\mu(\emptyset) = 0$;
- (2) μ is countably additive: for any collection of disjoint sets $\{A_n\}_{n \geq 1}$ in \mathcal{S} ,

$$\mu\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Evidently, if X is a Lebesgue measurable set and $\mathcal{S} = \mathfrak{L}(X)$, then λ is a measure on \mathcal{S} . But, for example, 2λ is also a measure on \mathcal{S} , and sometimes we may want to consider a “normalized” measure such as $\frac{1}{2}\lambda$ on $\mathfrak{L}([-1, 1])$.

While we mainly will be concerned with Lebesgue spaces, there are many situations where the main property of the space that is used is the countable additivity of the measure. So we introduce the notion of a *measure space*.

A **measure space** is defined to be a triple (X, \mathcal{S}, μ) where X is a nonempty set, \mathcal{S} is a σ -algebra in X and μ is a measure on \mathcal{S} . The elements of \mathcal{S} are sets that are said to be measurable with respect to \mathcal{S} or \mathcal{S} -measurable. A **probability space** is a measure space (X, \mathcal{S}, μ) such that $\mu(X) = 1$.

A measure space (X, \mathcal{S}, μ) is a **finite measure space** if $\mu(X) < \infty$ and it is **σ -finite** if there is a sequence of measurable sets A_n of finite measure such that $X = \bigcup_{n=1}^{\infty} A_n$. A σ -finite measure space may be of finite measure. Whenever we consider a measure space that has infinite measure we shall always assume it is σ -finite. Evidently, the real line with Lebesgue sets and Lebesgue measure is a σ -finite measure space, and the reader is asked to verify that all canonical Lebesgue measure spaces are σ -finite.

Another property that is enjoyed by Lebesgue measure is that of being *complete*. A measure space (X, \mathcal{S}, μ) is called **complete** if whenever $A \in \mathcal{S}$ and $\mu(A) = 0$, then for every $B \subset A$ we have $B \in \mathcal{S}$ (so $\mu(B) = 0$). In other words, in a complete measure space every null set (a set contained in a set of measure 0) is measurable.

We now define an important class of measure spaces. A **canonical nonatomic Lebesgue measure space** is a measure space $(X_0, \mathfrak{L}(X_0), \lambda)$ where X_0 is a (bounded or unbounded) interval in \mathbb{R} and λ is Lebesgue measure on $\mathfrak{L}(\mathbb{R})$.

The next important class of examples are the **canonical atomic spaces** that we now define. Let Z be any nonempty subset of \mathbb{Z} and let the σ -algebra \mathcal{S} be $\mathcal{P}(Z)$; if Z is a finite set let $\#(Z)$ denote its number of elements. The main examples that we consider are when

Z is $Z_n = \{0, \dots, n-1\}$, or \mathbb{Z} , or \mathbb{N}_0 . For each $k \in Z$ define

$$\nu(\{k\}) = \begin{cases} 1/\#(Z), & \text{if } Z \text{ is finite;} \\ 1, & \text{if } Z \text{ is infinite,} \end{cases}$$

$$\nu(\emptyset) = 0,$$

and extend ν to \mathcal{S} by $\nu(A) = \sum_{k \in A} \nu(\{k\})$. The reader should verify that ν is countably additive on \mathcal{S} . The points x_i are called the **atoms** of the space, and the elements of $\mathcal{P}(X)$ are the measurable sets of the space; every measurable set of the space is a finite union of atoms. The measure ν on Z_n can be thought of as modeling the tosses of a fair n -sided die. When $Z = \mathbb{Z}$ we call ν a **counting measure**.

A **canonical Lebesgue measure space** is defined to be a triple $(X, \mathcal{S}(X), \mu)$ such that

$$X = X_0 \sqcup Z,$$

where $(X_0, \mathfrak{L}(X_0), \lambda)$ is a canonical nonatomic Lebesgue space, and $(Z, \mathcal{P}(Z), \nu)$ is a canonical atomic Lebesgue space. The σ -algebra on X is given by $\mathcal{S}(X) = \mathfrak{L}(X_0) \sqcup \mathcal{P}(Z)$, and μ is defined by

$$\mu(A) = \begin{cases} \lambda(A), & \text{if } A \in \mathfrak{L}(X_0); \\ \nu(A), & \text{if } A \in \mathcal{P}(Z). \end{cases}$$

Question. Show that $\mathcal{S}(X) = \mathfrak{L}(X_0) \sqcup \mathcal{P}(Z)$ is a σ -algebra on X . In Exercise 2 the reader is asked to show that μ is a measure on $\mathcal{S}(X)$.

A *Lebesgue space* will be defined later to be any measure space *isomorphic* to a canonical Lebesgue measure space.

The following exercise and proposition are two examples of statements that only use the countable additivity property of the measure μ , and thus hold in the more general setting of measure spaces.

Question. Let (X, \mathcal{S}, μ) be a measure space. Show that if A, B are measurable sets with $A \subset B$, then $\mu(A) \leq \mu(B)$, and if $\mu(A) < \infty$, then $\mu(B \setminus A) = \mu(B) - \mu(A)$.

Proposition 2.5.2. *Let (X, \mathcal{S}, μ) be a measure space.*

- (1) *If $\{A_n\}_{n \geq 1}$ is a sequence of measurable sets in X that is increasing, i.e.,*

$$A_n \subset A_{n+1} \quad \text{for all } n \geq 1,$$

then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mu(A_n).$$

- (2) *If $\{B_n\}_{n \geq 1}$ is a sequence of measurable sets in X that is decreasing, i.e.,*

$$B_n \supset B_{n+1} \quad \text{for all } n \geq 1,$$

and $\mu(B_K) < \infty$ for some $K > 0$, then

$$\mu\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n \rightarrow \infty} \mu(B_n).$$

Proof. To prove part (1), let $A = \bigcup_{n=1}^{\infty} A_n$. If $\mu(A_n) = \infty$ for some $n > 0$, then $\mu(A) = \infty$ and $\mu(A_m) = \infty$ for all $m \geq n$, and we are done; so assume that $\mu(A_n) < \infty$ for all $n > 0$. Write $A_0 = \emptyset$ and observe that

$$A = \bigsqcup_{n=0}^{\infty} (A_{n+1} \setminus A_n).$$

The sets in the above union are disjoint as the A_n are increasing. Thus, by countable additivity,

$$\begin{aligned} \mu(A) &= \sum_{n=0}^{\infty} \mu(A_{n+1} \setminus A_n) = \lim_{N \rightarrow \infty} \sum_{n=0}^{\infty} \mu(A_{n+1} \setminus A_n) \\ &= \lim_{N \rightarrow \infty} \sum_{n=0}^N \mu(A_{n+1}) - \mu(A_n) = \lim_{N \rightarrow \infty} \mu(A_{N+1}), \end{aligned}$$

which completes the proof of part a).

To prove part (2) note that since $\bigcap_{n=1}^{\infty} B_n = \bigcap_{n=K}^{\infty} B_n$, after renaming the sets if necessary we may assume that $\mu(B_n) < \infty$ for all $n \geq 1$ (so $K = 1$). Write

$$C = \bigcap_{n=1}^{\infty} B_n.$$

Now note that

$$(2.4) \quad B_1 = C \sqcup \bigsqcup_{n=1}^{\infty} (B_n \setminus B_{n+1}).$$

Since the sets in (2.4) are disjoint, by countable additivity,

$$\mu(B_1) = \mu(C) + \sum_{n=1}^{\infty} \mu(B_n \setminus B_{n+1}).$$

Then we observe that

$$\begin{aligned} \sum_{n=1}^{\infty} \mu(B_n \setminus B_{n+1}) &= \lim_{N \rightarrow \infty} \sum_{n=1}^{N-1} (\mu(B_n) - \mu(B_{n+1})) \\ &= \mu(B_1) - \lim_{N \rightarrow \infty} \mu(B_N). \end{aligned}$$

Therefore,

$$\mu(B_1) - \mu(C) = \mu(B_1) - \lim_{N \rightarrow \infty} \mu(B_N),$$

so

$$\mu(C) = \lim_{N \rightarrow \infty} \mu(B_N).$$

This concludes the proof. \square

Exercises

- (1) Prove Proposition 2.5.1.
- (2) Let μ be as in the definition of a canonical Lebesgue space. Show that μ is a measure.
- (3) Show that Proposition 2.5.2, part b), does not hold without the assumption that $\mu(B_K) < \infty$ for some $K \geq 1$.
- (4) Let A and B be measurable sets such that $\mu(A) < \infty$. Let $\varepsilon > 0$. Show that $\mu(A \setminus B) < \varepsilon$ if and only if $\mu(A \cap B) > \mu(A) - \varepsilon$.
- (5) Let A and B be measurable sets such that $\mu(A) < \infty$. Let $\varepsilon > 0$. Show that if $\mu(A \triangle B) < \varepsilon$, then $\mu(A) - \varepsilon < \mu(B) < \mu(A) + \varepsilon$.
- (6) (Triangle Inequality) Let A, B, C be measurable sets. Show that

$$\mu(A \triangle B) \leq \mu(A \triangle C) + \mu(C \triangle B).$$

- (7) Show that if μ is a measure on a σ -algebra \mathcal{S} of some set X , then it must be countably subadditive, i.e., for any sets $A_n \in \mathcal{S}$, $\mu(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$.
- (8) Show that a canonical Lebesgue measure space is a σ -finite and complete measure space.
- (9) Let $X = \mathbb{R}$ and let \mathcal{S} be the collections of all subsets of X . For $A \subset X$ define $\mu(A)$ to be the number of elements in A if A is finite, and ∞ otherwise. Show that (X, \mathcal{S}, μ) is a measure space that is not σ -finite.
- (10) Let X be a Lebesgue measurable set. Given any two Lebesgue measurable sets A and B in X , define the relation $A \sim B$ when $\lambda(A \Delta B) = 0$. Show that \sim is an equivalent relation on the elements of $\mathfrak{L}(X)$.
- (11) Let $\{A_n\}, n \geq 1$, be a sequence of subsets of a set X . Define

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n,$$

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n.$$

Show that $\liminf_{n \rightarrow \infty} A_n$ consists of the sets of points $x \in X$ that are in A_n for all large n (we say that they are eventually in A_n), and that $\limsup_{n \rightarrow \infty} A_n$ consists of the sets of points $x \in X$ that are in infinitely many A_n .

- (12) (Borel–Cantelli) Let (X, \mathcal{S}, μ) be a probability space and let $\{A_n\}$ be a sequence of measurable sets. Show that if $\sum_{n=1}^{\infty} \mu(A_n) < \infty$, then $\mu(\limsup A_n) = 0$.
- (13) Let $[A]$ denote the equivalence class of the measurable set A under the equivalence relation of Exercise 10. Let $\mathfrak{L}(X)/\sim$ denote the set of equivalence classes, so $\mathfrak{L}(X)/\sim = \{[A] : A \in \mathfrak{L}(X)\}$. For any two equivalence classes $[A]$ and $[B]$ define $d([A], [B]) = \lambda(A \Delta B)$. Show that d is a metric on $\mathfrak{L}(X)/\sim$.

The following exercises develop the notion of *completion* of a measure space.

- (14) Let (X, \mathcal{S}, μ) be a measure space. Define the **completion** of \mathcal{S} with respect to μ to be the collection of sets \mathcal{S}_μ consisting of all sets $E \subset X$ such that there exist $A, B \in \mathcal{S}$ with

$$A \subset E \subset B \text{ and } \mu(B \setminus A) = 0.$$

Show that \mathcal{S}_μ is a σ -algebra containing \mathcal{S} .

- (15) Let (X, \mathcal{S}, μ) be a measure space and let \mathcal{S}_μ be as in Exercise 14. Define $\bar{\mu}$ on elements of \mathcal{S}_μ by $\bar{\mu}(E) = \mu(A)$ for any $A \in \mathcal{S}$ such that there is a $B \in \mathcal{S}$ with $A \subset E \subset B$ and $\mu(B \setminus A) = 0$. Show that the value of $\bar{\mu}$ on E is independent of A and B and therefore is a well-defined set function. Furthermore, show that $\bar{\mu}$ is a complete measure on \mathcal{S}_μ . We say that $(X, \mathcal{S}_\mu, \bar{\mu})$ is the (measure) completion of (X, \mathcal{S}, μ) .

2.6. The Borel Sigma-Algebra

This section introduces the notion of a collection of sets generating a σ -algebra and defines the σ -algebra of Borel sets, a σ -algebra properly contained in the σ -algebra of Lebesgue sets.

Given any two σ -algebras \mathcal{S}_1 and \mathcal{S}_2 on a nonempty set X , their intersection $\mathcal{S}_1 \cap \mathcal{S}_2$, defined by

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \{A : A \in \mathcal{S}_1 \text{ and } A \in \mathcal{S}_2\},$$

is also a σ -algebra. This follows from the simple fact that any collection of sets that belongs to the intersection also belongs to each σ -algebra, so the collection is closed under complements and countable unions; also the intersection contains at least one element, namely X . By a similar reasoning, only with minor changes to take care of the more complicated notation, the intersection of any (countable or not) collection of σ -algebras on a nonempty set X is a σ -algebra. Given any collection of subsets \mathcal{C} of a nonempty set X we define the **σ -algebra generated** by \mathcal{C} to be the intersection of all the σ -algebras containing \mathcal{C} ; this σ -algebra is denoted by $\sigma(\mathcal{C})$. Note that there is always at least one σ -algebra containing \mathcal{C} , namely $\mathcal{P}(X)$, the power set of X . It follows that $\sigma(\mathcal{C})$ is defined for any collection \mathcal{C} and it is a σ -algebra; it is characterized by the fact that if \mathcal{A} is any σ -algebra

containing \mathcal{C} , then $\sigma(\mathcal{C}) \subset \mathcal{A}$. The reader should verify that we have proved the following lemma.

Lemma 2.6.1. *Let X be a nonempty set and let \mathcal{C} be a collection of subsets of X . Then the σ -algebra generated by \mathcal{C} and denoted $\sigma(\mathcal{C})$ is the unique σ -algebra containing \mathcal{C} and such that for any σ -algebra \mathcal{A} containing \mathcal{C} it is the case that $\sigma(\mathcal{C}) \subset \mathcal{A}$.*

We are ready to define another of the important σ -algebras in analysis. The σ -algebra of **Borel sets** \mathcal{B} in \mathbb{R} is defined to be the σ -algebra generated by the open sets. Recall that \mathcal{G} stands for the collection of open sets, so

$$\mathcal{B} = \sigma(\mathcal{G}).$$

By definition, \mathcal{B} contains the open sets and the closed sets, and one might be tempted to think of the Borel sets as being obtained from the open sets by a countable number of unions, intersections and complements. However, the description of the Borel sets is subtler and the definition we have given of \mathcal{B} is nonconstructive; to specify all of its members one needs to make use of transfinite induction.

To give an idea of the complexity of the Borel sets we introduce some notation. If \mathcal{A} is a collection of sets, we let \mathcal{A}_δ denote the collection of countable intersections of sets from \mathcal{A} , and the collection of all countable unions of sets from \mathcal{A} is denoted by \mathcal{A}_σ . We let $\mathcal{A}_{\delta\sigma}$ denote $(\mathcal{A}_\delta)_\sigma$, etc. Let \mathcal{F} denote the closed sets (this comes from the French, *fermé*, for closed). Then we have the following proper inclusions for classes of Borel sets:

$$\begin{aligned} \mathcal{G} \subsetneq \mathcal{G}_\delta \subsetneq \mathcal{G}_{\delta\sigma} \subsetneq \mathcal{G}_{\delta\sigma\delta} \subsetneq \dots, \\ \mathcal{F} \subsetneq \mathcal{F}_\sigma \subsetneq \mathcal{F}_{\sigma\delta} \subsetneq \mathcal{F}_{\sigma\delta\sigma} \subsetneq \dots \end{aligned}$$

It can be shown that there are Borel sets that are not in the union of the classes above, and to obtain all the Borel sets we need to extend the classes above for all countable ordinals using transfinite induction. We do not do this here, as it is beyond the scope of this book.

All Borel sets are Lebesgue measurable, since the collection of Lebesgue measurable sets is a σ -algebra containing the open sets, and since the Borel σ -algebra is the smallest σ -algebra containing the

open sets it must be contained in the Lebesgue σ -algebra. Thus we have seen that $\mathcal{B} \subset \mathcal{L}$. Given a Borel set $X \subset \mathbb{R}$ we will use $\mathcal{B}(X)$ to denote the σ -algebra of Borel sets contained in X . By a similar reasoning, it follows that $\mathcal{B}(X) \subset \mathcal{L}(X)$

Furthermore, it can be shown that there are Lebesgue measurable sets that are not Borel sets. We can give the idea of this argument; it is based on showing that the cardinality of the collection of Lebesgue measurable sets is greater than the cardinality of the collection of Borel sets. Let \mathfrak{c} denote the cardinality of the continuum \mathbb{R} . The set of all subsets of \mathbb{R} has cardinality $2^{\mathfrak{c}}$, and is greater than \mathfrak{c} (by the Schroeder-Berstein theorem of set theory). Therefore there are at most $2^{\mathfrak{c}}$ Lebesgue measurable sets. But there exists an uncountable set K of measure zero, namely the Cantor set K defined in Section 2.2. All subsets of K must be of measure zero. Hence they are measurable. Since there are $2^{\mathfrak{c}}$ subsets of K it follows that there are at least $2^{\mathfrak{c}}$ Lebesgue measurable sets. Therefore there are $2^{\mathfrak{c}}$ Lebesgue measurable sets. Using transfinite induction it can be shown that the cardinality of the collection of Borel sets is \mathfrak{c} . (We just give an argument showing that the cardinality of the collection of open sets is \mathfrak{c} . Note that each open set can be written as a countable disjoint union of intervals with rational endpoints, which can be identified with the collection of all subsets of \mathbb{Q} , which has cardinality \mathfrak{c} .) Thus, there are Lebesgue sets that are not Borel. So we have that $\mathcal{B}(\mathbb{R}) \subsetneq \mathcal{L}(\mathbb{R})$. It also follows that the Borel σ -algebra does not contain all sets of measure zero.

As we have mentioned, there is no constructive way to describe the σ -algebra generated by a set. However, there is a useful approach to the generated σ -algebra in terms of monotone classes. A **monotone class** on a nonempty set X is a nonempty collection \mathcal{M} of subsets of X that is closed under countable increasing unions and countable decreasing intersections. As the power set of X is a monotone class, we can define the **monotone class generated by** a collection \mathcal{C} as the intersection of all monotone classes containing \mathcal{C} . Instead of considering the monotone class generated by an arbitrary collection, we shall put some additional structure on the generating collection. Define an **algebra** on a nonempty set X to be a nonempty collection

of subsets of X that is closed under complements and finite unions. It follows that it must be closed under finite intersections. Of course, any σ -algebra is an algebra, but the collection of all finite unions of intervals in $[0, 1]$ is an algebra that is not a σ -algebra. We start with the following elementary lemma.

Lemma 2.6.2. *Let X be a nonempty set. If a monotone class \mathcal{M} is an algebra, then it is a σ -algebra.*

Proof. It suffices to show that \mathcal{M} is closed under countable unions. Let $\{A_n\}$ be a sequence of sets in \mathcal{M} . Define $A'_n = \bigcup_{i=1}^n A_i$. Then the sequence $\{A'_n\}$ is monotone increasing, so its union is in \mathcal{M} , but as it has the same union as the sequence $\{A_n\}$, it follows that $\bigcup_{n=1}^{\infty} A_n$ is in \mathcal{M} . \square

We are ready to prove the theorem that characterizes the generated σ -algebra in terms of monotone classes. While this theorem is useful, its proof is somewhat mysterious. An extension of the theorem where “algebra” is replaced by “ring” appears in Exercise 2.7.12.

Theorem 2.6.3 (Monotone Class Theorem). *Let X be a nonempty set and let \mathcal{A} be an algebra of subsets of X . Then, the σ -algebra generated by \mathcal{A} , $\sigma(\mathcal{A})$, is equal to the monotone class generated by \mathcal{A} , denoted by \mathcal{M} .*

Proof. We show that $\sigma(\mathcal{A})$ is contained in \mathcal{M} . The other direction is simpler and left as an exercise. We observe that as \mathcal{M} contains \mathcal{A} , by Lemma 2.6.2, it suffices to show that \mathcal{M} is an algebra. For each $E \subset X$ define

$$\mathcal{M}(E) = \{A \subset X : A \cup E, A \setminus E, E \setminus A \in \mathcal{M}\}.$$

As \mathcal{M} is a monotone class, it readily follows that when the collection $\mathcal{M}(E)$ is nonempty, it is a monotone class. Also, if one takes E to be in the algebra \mathcal{A} , then $\mathcal{A} \subset \mathcal{M}(E)$. So the monotone class \mathcal{M} must be contained in $\mathcal{M}(E)$. Thus, for any A in \mathcal{M} and any E in \mathcal{A} , $A \in \mathcal{M}(E)$, which is equivalent to $E \in \mathcal{M}(A)$. So for any A in \mathcal{M} , $\mathcal{M} \subset \mathcal{M}(A)$. This, with the fact that X is in \mathcal{M} , means that \mathcal{M} is an algebra and completes the proof. \square

Exercises

- (1) Show that the intersection of any collection of σ -algebras is a σ -algebra.
- (2) Give an example of a \mathcal{G}_δ set that is neither open nor closed.
- (3) Show that every null set is contained in a Borel null set.
- (4) Let \mathcal{A} be a collection of subsets of a set X that contains X and is closed under the operation of set difference (i.e., if $A, B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$) and countable disjoint unions. Show that \mathcal{A} is a σ -algebra. (All that is needed is to show that \mathcal{A} is closed under countable unions.)
- (5) Show that the σ -algebra generated by the collection of all closed intervals with rational endpoints is equal to the Borel σ -algebra.
- (6) Show that the Borel σ -algebra is generated by \mathcal{F} , the collection of closed sets in \mathbb{R} .
- (7) Show that $(\mathbb{R}, \mathcal{L}, \lambda)$ is the completion (in the sense of Exercise 2.5.15) of $(\mathbb{R}, \mathcal{B}, \lambda)$.
- (8) Show that the set consisting of all open subsets of \mathbb{R} can be put in a one-to-one correspondence with \mathbb{R} .
- (9) Let X be a Borel subset of \mathbb{R} and let $\mathcal{B}(X)$ denote the σ -algebra of Borel sets contained in X . Show that $(X, \mathcal{B}(X), \lambda)$ is a measure space that is σ -finite but not complete. (You may use that there exists a one-to-one correspondence between the collection of Borel sets and \mathbb{R} .)
- (10) Show that in the proof of Theorem 2.6.3, \mathcal{M} is contained in $\sigma(\mathcal{A})$.

2.7. Approximation with Semi-rings

Intervals play an important role when approximating Lebesgue measurable sets. In the more general setting, semi-rings replace the collection of intervals. In this section we study notions of approximation in more detail and study approximating collections more general than the collection of all intervals.

When proving properties of measurable sets, often these properties are first shown for intervals or for finite unions of intervals, and then an approximation argument is used to extend the property to all measurable sets. In many cases, however, it will be convenient or necessary to consider a collection different from the collection of all intervals. For example, one may consider intervals with dyadic rational endpoints, or the intersection of intervals with a given measurable set, or perhaps a more general collection. These more general collections should be in some sense similar to the collection of intervals. We extract two basic properties from the collection of intervals: one captures how intervals behave under set-theoretic operations, and the other characterizes the approximation property. For the first one, note that

- the intersection of any two intervals is an interval, and
- the set difference of two intervals is a finite union of disjoint intervals.

These considerations will lead us to define the notion of a *semi-ring*. The approximation property is captured by the notion of a *sufficient semi-ring*.

On a first reading the reader may concentrate only on the definition of a semi-ring and sufficient semi-ring and the statement of Lemma 2.7.3. Here “sufficient semi-ring” may be replaced by the collection of all intervals or the collection of all dyadic intervals. In fact, the reader has already proven a version of this lemma in Exercise 2.4.3.

A **semi-ring** on a nonempty set X is a collection \mathcal{R} of subsets of X such that

- (1) \mathcal{R} is nonempty;
- (2) if $A, B \in \mathcal{R}$, then $A \cap B \in \mathcal{R}$;
- (3) if $A, B \in \mathcal{R}$, then

$$A \setminus B = \bigsqcup_{j=1}^n E_j,$$

where $E_j \in \mathcal{R}$ are disjoint.

A semi-ring must contain the empty set as $\emptyset = A \setminus A$ for any element $A \in \mathcal{R}$. Note that the collection of all intervals (or all bounded intervals) in \mathbb{R} forms a semi-ring. (This is clear since when I and J are any intervals, then $I \cap J$ is empty or an interval, and $I \setminus J$ is empty or a finite union of disjoint intervals, and intervals may be empty.) Similarly, one can verify that the collection of left-closed and right-open intervals is a semi-ring. Also, if \mathcal{R} is a semi-ring of subsets of a nonempty set X and $Y \subset X$, then the collection $\{A : A \cap Y, A \in \mathcal{R}\}$ is a semi-ring. Another example of a semi-ring is given by the collection of all (left-closed, right-open) dyadic intervals. For any set X , the collection of all its subsets is a semi-ring.

One reason semi-rings will be useful is the following proposition. It states that a set that is a countable union of elements of a semi-ring can be written as a countable disjoint union of elements of the semi-ring. (The reader should verify this for intervals.) The process in the proof of Proposition 2.7.1 will be called the process of “disjointifying” the sets A_n and will be useful in later chapters.

Proposition 2.7.1. *Let \mathcal{R} be a semi-ring. If $A = \bigcup_{n=1}^{\infty} A_n$, where $A_n \in \mathcal{R}$, then A can be written as*

$$A = \bigsqcup_{k=1}^{\infty} C_k,$$

where the sets $\{C_k\}$ are disjoint and are in the semi-ring \mathcal{R} .

Proof. Define a sequence of sets $\{B_n\}$ by $B_1 = A_1$, and for $n > 1$, $B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1})$. The sets $\{B_n\}$ are disjoint and

$$\bigcup_{n \geq 1} A_n = \bigsqcup_{n \geq 1} B_n.$$

However, the B_n 's need not be in \mathcal{R} . Now we show that each B_n in turn can be written as a finite union of disjoint elements of \mathcal{C} . First set $C_1 = B_1$. Next note that as $B_2 = A_2 \setminus A_1$, by property (3) of a semi-ring, B_2 can be written as a finite union of disjoint elements of \mathcal{R} ; call them C_2, \dots, C_{k_1} .

For $n \geq 3$, observe that

$$A_n \setminus (A_1 \cup \dots \cup A_{n-1}) = (A_n \setminus A_1) \cap (A_n \setminus A_2) \cap \dots \cap (A_n \setminus A_{n-1}).$$

Furthermore, each $(A_n \setminus A_i)$ can be written as a finite union of disjoint sets in \mathcal{R} , so

$$(A_n \setminus A_i) = \bigsqcup_{k=1}^{K_{n,i}} E_k^{n,i},$$

for some $E_k^{n,i} \in \mathcal{R}$. But \mathcal{R} is closed under finite intersections and by taking all possible intersections of the sets $E_k^{n,i}$ one can write each set B_n as a finite union of disjoint sets in \mathcal{R} . We show this for the case when $n = 3$ and the general case follows by induction. Write

$$A_3 \setminus A_1 = \bigsqcup_{k=1}^{K_{3,1}} E_k^{3,1} \quad \text{and} \quad A_3 \setminus A_2 = \bigsqcup_{\ell=1}^{K_{3,2}} E_\ell^{3,2}.$$

Then let

$$F_{k,\ell} = E_k^{3,1} \cap E_\ell^{3,2},$$

for $k = 1, \dots, K_{3,1}$, $\ell = 1, \dots, K_{3,2}$. Then the family of sets $\{F_{k,\ell}\}$, for $k = 1, \dots, K_{3,1}$, $\ell = 1, \dots, K_{3,2}$, is disjoint and are all elements of \mathcal{R} . It follows that

$$B_3 = (A_3 \setminus A_1) \cap (A_3 \setminus A_2) = \bigsqcup_{k,\ell} F_{k,\ell},$$

a disjoint finite union of elements of the semi-ring \mathcal{R} , which we rename $C_{k_1+1}, \dots, C_{k_2}$.

This yields a collection of disjoint sets $\{C_k\}$ in \mathcal{R} whose union is A . □

Let (X, \mathcal{S}, μ) be a measure space (in most applications, a canonical Lebesgue measure space). A semi-ring \mathcal{C} of measurable subsets of X of finite measure is said to be a **sufficient semi-ring** for (X, \mathcal{S}, μ) if it satisfies the following approximation property:

For every $A \subset \mathcal{S}$,

$$(2.5) \quad \mu(A) = \inf \left\{ \sum_{j=1}^{\infty} \mu(I_j) : A \subset \bigcup_{j=1}^{\infty} I_j \text{ and } I_j \in \mathcal{C} \text{ for } j \geq 1 \right\}.$$

The definition of a sufficient semi-ring is of interest mainly in the case of nonatomic spaces. The arguments in the atomic case are rather straightforward and will be left to the reader. We have seen,

for example, that the collection of all intervals, the collection of all intervals with rational endpoints and the collection of (left-closed, right-open) dyadic intervals are all sufficient semi-rings for $(\mathbb{R}, \mathfrak{L}, \lambda)$. Also, for any measurable set $X \subset \mathbb{R}$, any of these collections of intervals intersected with X forms a sufficient semi-ring for $(X, \mathfrak{L}, \lambda)$ (see Exercise 2). The following lemmas demonstrate why the class of sufficient semi-rings is interesting, and enable us to prove Theorem 3.4.1. The first lemma shows how one can approximate measurable sets up to measure zero with elements from a sufficient semi-ring, but uses countably many of them. The second lemma uses only finitely many elements of the sufficient semi-ring, but in this case the approximation is only “up to ε .”

Lemma 2.7.2. *Let (X, \mathcal{S}, μ) be a measure space with a sufficient semi-ring \mathcal{C} . Then for any $A \in \mathcal{S}$ with $\mu(A) < \infty$ there exists a set $H = H(A)$, of the form*

$$H = \bigcap_{n=1}^{\infty} H_n,$$

and such that

- (1) $H_1 \supset H_2 \supset \cdots \supset H_n \supset \cdots \supset H \supset A$;
- (2) $\mu(H_n) < \infty$;
- (3) each H_n is a countable disjoint union of elements of \mathcal{C} ; and
- (4) $\mu(H \setminus A) = 0$.

Proof. By the approximation property (2.5) of \mathcal{C} , for any $\varepsilon > 0$ there is a set $H(\varepsilon) = \bigcup_{j=1}^{\infty} I_j$, with $I_j \in \mathcal{C}$, such that

$$A \subset H(\varepsilon) \text{ and } \mu(H(\varepsilon) \setminus A) < \varepsilon.$$

Write

$$H_n = H(1) \cap H(1/2) \cap \cdots \cap H(1/n).$$

Since \mathcal{C} is closed under finite intersections, one can verify that each H_n is a countable union of elements of \mathcal{C} , and by Proposition 2.7.1 we can write this union as a countable disjoint union of elements of \mathcal{C} . Furthermore, by construction the sets H_n are decreasing (i.e.,

$H_n \subset H_m$ for $n > m$), contain A , and each H_n is of finite measure. Let

$$H = \bigcap_{n=1}^{\infty} H_n.$$

Then H has the required form and

$$\mu(H \setminus A) \leq \mu(H_n \setminus A) < 1/n,$$

for all $n \geq 1$. Therefore $\mu(H \setminus A) = 0$, so $\mu(A) = \mu(H)$. \square

We can say that the set H of Lemma 2.7.2 is in $\mathcal{C}_{\sigma\delta}$.

The following lemma will have several applications in later chapters. It is the first of Littlewood's Three Principles, which says that "every (Lebesgue) measurable set is nearly a finite union of intervals."

Lemma 2.7.3. *Let (X, \mathcal{S}, μ) be a measure space with a sufficient semi-ring \mathcal{C} . Let A be a measurable set, $\mu(A) < \infty$, and let $\varepsilon > 0$. Then there exists a set H^* that is a finite union of disjoint elements of \mathcal{C} such that*

$$\mu(A \Delta H^*) < \varepsilon.$$

Proof. As \mathcal{C} is a sufficient semi-ring and $\mu(A) < \infty$, for $\varepsilon > 0$, there exists a set $H[\varepsilon] = \bigcup_{j=1}^{\infty} I_j \supset A$ with $I_j \in \mathcal{C}$ and $\mu(H[\varepsilon]) < \mu(A) + \varepsilon/2$, so $\mu(H[\varepsilon] \setminus A) < \varepsilon/2$. By Proposition 2.5.2,

$$\lim_{n \rightarrow \infty} \mu\left(\bigcup_{j=1}^n I_j\right) = \mu\left(\bigcup_{j=1}^{\infty} I_j\right),$$

so there exists $N > 1$ such that

$$0 \leq \mu(H[\varepsilon]) - \mu\left(\bigcup_{j=1}^N I_j\right) < \varepsilon/2.$$

Let $H^* = \bigcup_{j=1}^N I_j$. Then

$$\begin{aligned} \mu(H^* \Delta A) &= \mu(H^* \setminus A) + \mu(A \setminus H^*) \\ &\leq \mu(H[\varepsilon] \setminus A) + \mu(H[\varepsilon] \setminus H^*) \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Finally, by Proposition 2.7.1, we may assume that H^* is a disjoint union of elements of \mathcal{C} . \square

We end this section with another characterization of approximation. First note that in Lemma 2.7.3, the set H^* is a finite union of elements of the semi-ring; so it will be useful to consider collections closed under finite unions. A semi-ring that is closed under finite unions is called a **ring**. In the case when X is of finite measure, most of the rings that we consider also contain X , so they have the additional structure of an algebra (an algebra is just a ring that contains X). The reader should keep the following examples in mind. The typical semi-ring is the collection of all bounded intervals in \mathbb{R} (or all left-closed, right-open bounded intervals in \mathbb{R}). The typical ring is the collection of all finite unions of such intervals. The typical algebra is the collection of all finite unions of subintervals of a bounded interval (or all left-closed, right-open subintervals of an interval of the form $[a, b)$). Refer to the exercises at the end of this section for other characterizations of rings and algebras.

Let (X, \mathcal{S}, μ) be a σ -finite measure space. (In most cases, X will be of finite measure.) A ring \mathcal{R} is said to **generate mod 0** the σ -algebra \mathcal{S} if the σ -algebra $\sigma(\mathcal{R})$ generated by \mathcal{R} satisfies the property: for all $A \in \mathcal{S}$, $\mu(A) < \infty$, there exists $E \in \sigma(\mathcal{R})$ such that $\mu(A \Delta E) = 0$. Evidently, the ring of finite unions of intervals generates mod 0 the Lebesgue sets. The following lemma gives a characterization of the notion of generation mod 0. A ring that generates mod 0 may also be called a **dense ring**. A **dense algebra** is defined similarly.

Lemma 2.7.4. *Let (X, \mathcal{S}, μ) be a finite measure space. Let \mathcal{A} be an algebra in X . Then \mathcal{A} generates \mathcal{S} mod 0 if and only if for any $A \in \mathcal{S}$ and any $\varepsilon > 0$ there exists $E \in \mathcal{A}$ such that $\mu(A \Delta E) < \varepsilon$.*

Proof. Suppose \mathcal{A} generates \mathcal{S} mod 0, so $\sigma(\mathcal{A})$ is equal to \mathcal{S} mod 0. Define the collection of sets

$$\mathcal{C} = \{A \in \sigma(\mathcal{A}) : \text{for } \varepsilon > 0 \text{ there exists } E \in \mathcal{A} \text{ with } \mu(A \Delta E) < \varepsilon\}.$$

Clearly, \mathcal{C} contains \mathcal{A} . As $\mu(A^c \Delta E^c) = \mu(A \Delta E)$, it follows that \mathcal{C} is closed under complements. With a bit more work one verifies that \mathcal{C} is closed under countable unions (see Exercise 8). Therefore \mathcal{C} is a σ -algebra. This implies that $\sigma(\mathcal{A})$ is contained in \mathcal{C} , completing the proof of this direction.

Now suppose that for any $A \in \mathcal{S}$ and any $\varepsilon > 0$ there exists $E \in \mathcal{A}$ such that $\mu(A \Delta E) < \varepsilon$. For each $n \geq 1$ choose a set $E_n \in \mathcal{A}$ such that

$$\mu(A \Delta E_n) < \frac{\varepsilon}{2^{n+1}}.$$

Then for each $k \geq 1$, $\mu(A \Delta \bigcup_{n=k}^{\infty} E_n) < \varepsilon$. So

$$\mu(A \Delta \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} E_n) = 0.$$

Then $F = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} E_n \in \sigma(\mathcal{A})$ and $\mu(A \Delta F) = 0$. \square

Corollary 2.7.5. *Let (X, \mathcal{S}, μ) be a finite measure space. Let \mathcal{C} be a semi-ring in \mathcal{S} containing X . Then \mathcal{C} is a sufficient semi-ring if and only if the algebra consisting of all finite unions of elements of \mathcal{C} is dense.*

We conclude with some alternative notation to express the notions of approximation. We think of two measurable sets A and B as being “ ε -close” (for some $\varepsilon > 0$) when $\lambda(A \Delta B) < \varepsilon$. We state some properties of this “distance.”

Proposition 2.7.6. *Let (X, \mathcal{S}, μ) be a probability measure space. For any measurable sets A and B , define*

$$D(A, B) = \mu(A \Delta B).$$

Then

- (1) $D(A, B) = 0$ if and only if $A = B \bmod \mu$;
- (2) $D(A, B) = D(B, A)$;
- (3) $D(A, B) \leq D(A, C) + D(C, B)$, for any measurable set C ;
- (4) if $D(A, B) < \varepsilon$, then $|\mu(A) - \mu(B)| < \varepsilon$.

It follows that D is a pseudo-metric; it is not a metric as it may happen that $A \neq B$ but $D(A, B) = 0$ (in fact, for any null set N , $D(N, \emptyset) = 0$); but D obeys the other properties of a metric. Also, using this notation it follows that if \mathcal{C} is a sufficient ring, then for all measurable sets A and any $\varepsilon > 0$ there exists $C \in \mathcal{C}$ so that $D(B, C) < \varepsilon$. When working with probability spaces, most rings that one studies already contain X . A ring that contains X is an algebra

(see Exercise 5). An important algebra in $[0, 1]$ is the collection of all finite unions of dyadic intervals.

Exercises

- (1) Show that if \mathcal{C} is a semi-ring of subsets of a nonempty set X and $\emptyset \neq Y \subset X$, then the collection $\{A \cap Y : A \in \mathcal{C}\}$, if nonempty, is a semi-ring of subsets of Y . Show a similar property for the case when \mathcal{C} is a ring.
- (2) Let $(X, \mathfrak{L}, \lambda)$ be a canonical Lebesgue measure space and \mathcal{C} a sufficient semi-ring. Show that for any nonempty measurable set $X_0 \subset X$, $\mathcal{C} \cap X_0 = \{C \cap X_0 : C \in \mathcal{C}\}$ is a sufficient semi-ring for $(X_0, \mathfrak{L}(X_0), \lambda)$.
- (3) Show that a ring is closed under symmetric differences and finite intersections.
- (4) Let \mathcal{R} be a nonempty collection of subsets of a set X such that for all $A, B \in \mathcal{R}$, $A \cup B \in \mathcal{R}$ and $A \setminus B \in \mathcal{R}$. Show that \mathcal{R} is a ring.
- (5) Let \mathcal{A} be a nonempty collection of subsets of a set X . Show that \mathcal{A} is an algebra if and only if it is closed under complements and finite intersections. Give an example of a ring that is not an algebra. Give an example of an algebra that is not a σ -algebra.
- (6) Show that if \mathcal{C} is a collection of sets, then the intersection of all rings containing \mathcal{C} is a ring, called the **ring generated by \mathcal{C}** and denoted by $r(\mathcal{C})$.
- (7) Show that if \mathcal{R} is a semi-ring, then the ring $r(\mathcal{R})$ generated by \mathcal{R} is obtained by taking all finite unions of disjoint elements from \mathcal{R} .
- (8) Let \mathcal{C} be defined as in the proof of Lemma 2.7.4. Show that \mathcal{C} is closed under countable unions.
- (9) Prove Proposition 2.7.6.
- (10) Prove Corollary 2.7.5.
- * (11) Let X be a Lebesgue measurable set in \mathbb{R} and let d be the metric defined on the set $\mathfrak{L}(X)/\sim$ in Exercise 2.5.13. Let \mathcal{C} be a sufficient semi-ring in $(X, \mathfrak{L}(X), \lambda)$ and let $r(\mathcal{C})$ be

the ring generated by \mathcal{C} . Show that in the metric space $(\mathfrak{L}(X)/\sim, d)$ the collection $\{[H] : H \in \mathcal{R}\}$ is a dense set.

- (12) Let X be a nonempty set and let \mathcal{R} be a ring of subsets of X . Define a σ -ring to be a ring that is closed under countable unions and countable intersections. Show that the smallest σ -ring containing \mathcal{R} is equal to the smallest monotone class containing \mathcal{R} .

2.8. Measures from Outer Measures

Lebesgue measure can be extended in a natural way from \mathbb{R} to \mathbb{R}^d for any integer $d > 1$. Lebesgue measure for $d = 2$ should generalize the notion of area, and for $d = 3$ the notion of volume. The idea is to replace intervals, in the definition of Lebesgue outer measure in the line, by Cartesian products of intervals in the definition of outer measure in \mathbb{R}^d . Rather than developing d -dimensional Lebesgue measure in a way similar to our construction of Lebesgue measure, we instead introduce a method that works in a more general context. This method is based on Carathéodory's condition for measurability.

Before presenting the main definitions we introduce some notation. A **set function** is a function μ , defined on some collection \mathcal{C} of subsets of a nonempty set X with $\emptyset \in \mathcal{C}$, such that μ has values in $[-\infty, \infty]$ (all the set functions we consider will have values in $[0, \infty]$). A set function $\mu : \mathcal{C} \rightarrow [0, \infty]$ is said to be **finitely additive** on \mathcal{C} if $\mu^*(\emptyset) = 0$ and for any disjoint sets $A_i, 1 \leq i \leq n$, in \mathcal{C} such that $\bigsqcup_{i=1}^n A_i \in \mathcal{C}$, it is the case that

$$\mu^*\left(\bigsqcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu^*(A_i).$$

A set function $\mu : \mathcal{C} \rightarrow [0, \infty]$ is **countably additive** on \mathcal{C} if $\mu^*(\emptyset) = 0$ and for any sequence of disjoint sets $A_i, i \geq 1$, in \mathcal{C} such that $\bigsqcup_{i=1}^{\infty} A_i \in \mathcal{C}$, it is the case that

$$\mu^*\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu^*(A_i).$$

Certainly, Lebesgue measure on the semi-ring of bounded subintervals of \mathbb{R} is a countably additive set function. A set function that is finitely

additive but not countably additive is given in Exercise 1. We will mainly be interested in finite additivity as an intermediate property that is verified before eventually showing that the set function under consideration is countably additive.

Given a nonempty set X , an **outer measure** μ^* is a set function defined on all subsets of X and with values in $[0, \infty]$ satisfying the following properties:

- (1) $\mu^*(\emptyset) = 0$;
- (2) μ^* is **monotone**: if $A \subset B$, then $\mu^*(A) \leq \mu^*(B)$;
- (3) μ^* is **countably subadditive**: $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$.

The first example we have of an outer measure is Lebesgue outer measure. Any measure defined on all subsets of a finite set is also an outer measure.

As we have seen, one of the most important properties to prove in our development of Lebesgue measure was countable additivity. Part of the proof involved choosing the right class of sets on which the outer measure is a measure, in this case the σ -algebra of Lebesgue measurable sets. Carathéodory introduced a remarkable property that does precisely this for any outer measure. A set A is said to be μ^* -**measurable** if for any set C ,

$$\mu^*(C) = \mu^*(A \cap C) + \mu^*(A^c \cap C).$$

This is the same as saying that a set A is μ^* -measurable if and only if for any set C_1 contained in A and any set C_2 contained in its complement A^c ,

$$(2.6) \quad \mu^*(C_1 \sqcup C_2) = \mu^*(C_1) + \mu^*(C_2).$$

(This follows by letting $C = C_1 \cup C_2$ in the definition.) In other words, A is μ^* -measurable if and only if A partitions the sets in X so that μ^* is additive on the disjoint union of a set contained in A with a set contained in its complement. While this characterization gives an idea of the condition, its true merit lies in the fact that Theorem 2.8.1 holds.

Question. Show that any set A with $\mu^*(A) = 0$ is μ^* -measurable.

Theorem 2.8.1. *Let X be a nonempty set and let μ^* be an outer measure on X . If $\mathcal{S}(\mu^*)$ denotes the μ^* -measurable sets in X , then $\mathcal{S}(\mu^*)$ is a σ -algebra and μ^* restricted to $\mathcal{S}(\mu^*)$ is a measure.*

Proof. It follows immediately from the definition that $\mathcal{S}(\mu^*)$ contains the empty set and is closed under complements. We now show that $\mathcal{S}(\mu^*)$ is closed under finite unions. First we show that this is the case for finite unions of disjoint sets. So let $A, B \in \mathcal{S}(\mu^*)$ with $A \cap B = \emptyset$. Let $C_1 \subset A \sqcup B$ and $C_2 \subset (A \sqcup B)^c$. Write $C_{1,1} = C_1 \cap A$ and $C_{1,2} = C_1 \cap B$. As A is μ^* -measurable, $C_{1,1} \subset A$, and $C_{1,2} \subset A^c$,

$$\mu^*(C_1) = \mu^*(C_{1,1} \sqcup C_{1,2}) = \mu^*(C_{1,1}) + \mu^*(C_{1,2}).$$

Similarly, using that B is μ^* -measurable,

$$\mu^*(C_{1,2} \sqcup C_2) = \mu^*(C_{1,2}) + \mu^*(C_2).$$

Therefore, again using that A is μ^* -measurable and then applying the previous results,

$$\begin{aligned} \mu^*(C_1 \sqcup C_2) &= \mu^*(C_{1,1} \sqcup (C_{1,2} \sqcup C_2)) = \mu^*(C_{1,1}) + \mu^*(C_{1,2} \sqcup C_2) \\ &= \mu^*(C_{1,1}) + \mu^*(C_{1,2}) + \mu^*(C_2) \\ &= \mu^*(C_{1,1} \sqcup C_{1,2}) + \mu^*(C_2) \\ &= \mu^*(C_1) + \mu^*(C_2). \end{aligned}$$

Therefore, $A \sqcup B$ is in $\mathcal{S}(\mu^*)$.

Next we show in a similar way that if $A, B \in \mathcal{S}(\mu^*)$, then $A \cap B \in \mathcal{S}(\mu^*)$. In fact, let $C_1 \subset A \cap B$ and $C_2 \subset A^c \cup B^c$. Let $C_{2,1} = C_2 \cap A$ and $C_{2,2} = C_2 \cap B$. Then

$$\begin{aligned} \mu^*(C_1 \sqcup C_2) &= \mu^*((C_1 \sqcup (C_{2,1} \sqcup C_2)) = \mu^*(C_1 \sqcup C_{2,1}) + \mu^*(C_{2,2}) \\ &= \mu^*(C_1) + \mu^*(C_{2,1}) + \mu^*(C_{2,2}) \\ &= \mu^*(C_1) + \mu^*(C_{2,1} \sqcup C_{2,2}) \\ &= \mu^*(C_1) + \mu^*(C_2). \end{aligned}$$

So $\mathcal{S}(\mu^*)$ is closed under finite intersections. Since $\mathcal{S}(\mu^*)$ is closed under complements, finite disjoint unions and finite intersections, then it is closed under finite unions.

Before considering countably many sets we observe that μ^* is finitely additive on μ^* -measurable sets. In fact if A and B are μ^* -measurable and disjoint, as $A \subset A$ and $B \subset A^c$, then $\mu^*(A \sqcup B) = \mu^*(A) + \mu^*(B)$.

It remains to show that $\mathcal{S}(\mu^*)$ is closed under countable unions and that μ^* is countably additive. Let $\{A_i\}_{i \geq 1}$ be a sequence of sets in $\mathcal{S}(\mu^*)$. We wish to verify that $\bigcup_i A_i$ is μ^* -measurable. Since we now know that $\mathcal{S}(\mu^*)$ is an algebra, hence a semi-ring, Proposition 2.7.1 implies that we may assume that the sets A_i are disjoint. Let

$$C_1 \subset \bigsqcup_{i=1}^{\infty} A_i, \quad C_2 \subset \left(\bigsqcup_{i=1}^{\infty} A_i \right)^c,$$

and set

$$C_{1,n} = C_1 \cap \bigsqcup_{i=1}^n A_i.$$

As $C_2 \subset (\bigsqcup_{i=1}^n A_i)^c$, $\mu^*(C_{1,n} \sqcup C_2) = \mu^*(C_{1,n}) + \mu^*(C_2)$. Therefore,

$$\begin{aligned} \mu^*(C_1 \sqcup C_2) &\geq \lim_{n \rightarrow \infty} \mu^*(C_{1,n} \sqcup C_2) \\ &= \lim_{n \rightarrow \infty} \mu^*(C_{1,n}) + \mu^*(C_2). \end{aligned}$$

But,

$$\begin{aligned} \mu^*(C_1) &= \mu^*\left(\bigsqcup_{i=1}^{\infty} A_i \cap C_1\right) \leq \sum_{i=1}^{\infty} \mu^*(A_i \cap C_1) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu^*(A_i \cap C_1) \\ &= \lim_{n \rightarrow \infty} \mu^*\left(\bigsqcup_{i=1}^n A_i \cap C_1\right) \\ &= \lim_{n \rightarrow \infty} \mu^*(C_{1,n}). \end{aligned}$$

Therefore, $\mu^*(C_1 \sqcup C_2) \geq \mu^*(C_1) + \mu^*(C_2)$. This shows that $\bigsqcup_{i=1}^{\infty} A_i \in \mathcal{S}(\mu^*)$.

The fact that μ^* is countably additive follows an argument we have already seen:

$$\begin{aligned}\mu^* \left(\bigsqcup_{i=1}^{\infty} A_i \right) &\geq \lim_{n \rightarrow \infty} \mu^* \left(\bigsqcup_{i=1}^n A_i \right) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu^*(A_i) = \sum_{i=1}^{\infty} \mu^*(A_i).\end{aligned}$$

□

Application: A Construction of Lebesgue measure on \mathbb{R}^d .

Recall that the **Cartesian product** of d sets A_1, \dots, A_d is defined to be

$$A_1 \times \dots \times A_d = \{(x_1, \dots, x_d) : x_i \in A_i, \text{ for } i = 1, \dots, d\}.$$

A **d -rectangle** or **rectangle** I in \mathbb{R}^d is defined to be a set of the form

$$I = I_1 \times \dots \times I_d,$$

where $I_j, j \in \{1, \dots, d\}$, are bounded intervals in \mathbb{R} .

Define a set function on the collection of d -rectangles by

$$|I|_{(d)} = |I_1| \times \dots \times |I_d|,$$

called the **d -volume**. Clearly, for $d = 1$ this corresponds to the length on an interval, for $d = 2$ to the area of a square and for $d = 3$ to the volume of a 3-rectangle.

Define the **d -dimensional Lebesgue outer measure** in \mathbb{R}^d by (2.7)

$$\lambda_{(d)}^*(A) = \inf \left\{ \sum_{j=1}^{\infty} |I_j|_d : A \subset \bigcup_{j=1}^{\infty} I_j, \text{ where } I_j \text{ are } d\text{-rectangles} \right\}.$$

By Theorem 2.8.1 this outer measure is a measure when restricted to the σ -algebra $S(\lambda_{(d)}^*)$. It remains to show that d -rectangles are $\lambda_{(d)}^*$ -measurable. This will imply that the Borel sets in \mathbb{R}^d are $\lambda_{(d)}^*$ -measurable, so $\lambda_{(d)}^*$ restricted to the Borel sets is a measure, giving a construction of Lebesgue measure on the Borel σ -algebra of \mathbb{R}^d . For a proof that the Borel sets are $\lambda_{(1)}^*$ -measurable in the 1-dimensional case, that easily extends to the d -dimensional case. For a complete

proof the reader may consult [14, Proposition 1.3.5]. Another construction is given later as an application of the following theorem.

One can use the notion of outer measure to construct more general measures by extending a finitely additive set function defined on a semi-ring of sets. When working with semi-rings we shall make the assumption that the set X can be written as a countable union of elements of the semi-ring \mathcal{R} . While some theorems can be proved without this assumption, all the cases we shall be interested in satisfy the assumption.

Theorem 2.8.2 (Carathéodory Extension Theorem). *Let X be a nonempty set and let \mathcal{R} be a semi-ring of subsets of X such that X can be written as a countable union of elements of \mathcal{R} . Let μ be a countably additive set function on \mathcal{R} . Then μ extends to a measure defined on the σ -algebra $\sigma(\mathcal{R})$ generated by \mathcal{R} .*

Proof. We first outline the structure of the proof. We start by constructing an outer measure μ^* from μ . Then we show that the elements of \mathcal{R} are μ^* -measurable, which implies that the σ -algebra generated by \mathcal{R} is contained in $\mathcal{S}(\mu^*)$. Thus μ^* defines a measure on $\sigma(\mathcal{R})$, which we also denote by μ .

Start by defining μ^* for any set $A \subset X$, by

$$\mu^*(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu(E_i) : A \subset \bigcup_{i=1}^{\infty} E_i, E_i \in \mathcal{R} \right\}.$$

We now show that μ^* is countably subadditive. This is similar to the corresponding proof for the case of Lebesgue outer measure.

Let $A = \bigcup_{i=1}^{\infty} A_i$. We may assume that $\mu(A_i) < \infty$ for $i \in \mathbb{N}$. Let $\varepsilon > 0$. Then for each i there exist sets $E_{i,j} \in \mathcal{R}, j \geq 1$, such that $A_i \subset \bigcup_{j=1}^{\infty} E_{i,j}$, and

$$\sum_{j=1}^{\infty} \mu(E_{i,j}) < \mu^*(A_i) + \frac{\varepsilon}{2^i}.$$

Then

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu(E_{i,j}) \leq \sum_{i=1}^{\infty} \mu^*(A_i) + \varepsilon.$$

As $A \subset \bigcup_i \bigcup_j E_{i,j}$, $\mu^*(A) \leq \sum_{i=1}^{\infty} \mu^*(A_i) + \varepsilon$. Letting $\varepsilon \rightarrow 0$ completes the argument. As the other properties of an outer measure are easy to verify we conclude that μ^* is an outer measure.

Next, we show that the elements of \mathcal{R} are μ^* -measurable. Let $E \in \mathcal{R}$ and suppose that $C_1 \subset E$ and $C_2 \subset E^c$. If $\mu^*(C_1) = \infty$ or $\mu^*(C_2) = \infty$, then $\mu^*(C_1 \sqcup C_2) = \infty$, and (2.6) is trivially verified. So suppose that $\mu^*(C_1 \sqcup C_2) < \infty$. For $\varepsilon > 0$ there exist $K_i \in \mathcal{R}$ such that $C_1 \sqcup C_2 \subset \bigsqcup_{i=1}^{\infty} K_i$ and

$$\mu^*(C_1 \sqcup C_2) > \left[\sum_{i=1}^{\infty} \mu(K_i) \right] - \varepsilon.$$

We can write $K_i = (K_i \cap E) \sqcup (K_i \cap E^c)$ and, as \mathcal{R} is a semi-ring,

$$K_i \cap E^c = K_i \setminus E = \bigsqcup_{j=1}^{n_i} F_{i,j},$$

for some $F_{i,j} \in \mathcal{R}$. Thus,

$$\begin{aligned} \mu^*(C_1 \sqcup C_2) &> \left[\sum_{i=1}^{\infty} \mu(K_i \cap E \sqcup \bigsqcup_{j=1}^{n_i} F_{i,j}) \right] - \varepsilon \\ &= \left[\sum_{i=1}^{\infty} \mu(K_i \cap E) \right] + \left[\sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \mu(F_{i,j}) \right] - \varepsilon \\ &\geq \mu^*(C_1) + \mu^*(C_2) - \varepsilon. \end{aligned}$$

This shows that the elements of \mathcal{R} are μ^* -measurable. Therefore $\sigma(\mathcal{R})$ is contained in $\mathcal{S}(\mu^*)$.

We observe that μ^* restricted to \mathcal{R} agrees with μ . Clearly, as $E \in \mathcal{R}$ covers itself, $\mu^*(E) \leq \mu(E)$. Then Exercise 2 implies that $\mu(E) \leq \mu^*(E)$. \square

The hypothesis of Theorem 2.8.2 can be relaxed slightly when combined with Exercise 5, whose proof uses ideas already discussed in this section.

Sometimes the Carathéodory extension theorem is stated in the following form. The first part is an immediate consequence of Theorem 2.8.2.

Theorem 2.8.3. *Let X be a nonempty set and let \mathcal{A} be an algebra of subsets of X . If μ is a countably additive set function on \mathcal{A} , then it extends to a measure on the σ -algebra generated by \mathcal{A} . If μ is σ -finite, then the extension is unique.*

Proof. The existence of the extension follows from Theorem 2.8.2. We show uniqueness of the extension. This is a standard argument for uniqueness and uses the monotone class theorem. We show this in the case when μ is finite and leave the general case to the reader. (For the general case the reader should consider $X = \bigcup_{i=1}^{\infty} K_i$, with $\mu(K_i) < \infty, K_i \in \mathcal{A}$.) Let ν be any other measure on $\sigma(\mathcal{A})$ that agrees with μ on all elements of \mathcal{R} . Form the set

$$\mathcal{C} = \{A \in \sigma(\mathcal{R}) : \mu(A) = \nu(A)\}.$$

Evidently, \mathcal{C} contains \mathcal{A} . If we show that \mathcal{C} is a monotone class, the monotone class theorem implies that \mathcal{C} contains $\sigma(\mathcal{A})$, completing the proof. But the fact that \mathcal{C} is closed under monotone unions and intersections, and is a monotone class, follows from Proposition 2.5.2. \square

We now discuss briefly that it can be shown that a countably additive set function on a semi-ring extends to the generated ring. We will typically not need this result, as we have shown in Theorem 2.8.2 a countably additive set function only needs to be defined on a semi-ring to have an extension to the generated σ -algebra, but is used for uniqueness of the extension as seen below. Recall that $r(\mathcal{R})$ denotes the ring generated by a semi-ring \mathcal{R} and consists of all finite unions of elements of \mathcal{R} (Exercise 2.7.4). We simply give a brief outline of the proof; for a complete proof see [68, Theorem 3.5].

Proposition 2.8.4. *Let \mathcal{R} be a semi-ring on a set X . If μ is a measure on \mathcal{R} , then it has a unique extension to $r(\mathcal{R})$.*

Proof. Let μ also denote the extension of μ . It has a natural definition on $r(\mathcal{R})$. Let $A \in r(\mathcal{R})$. Then A can be written as $A = \bigsqcup_{i=1}^n K_i$ for some $K_i \in \mathcal{R}$ and $n > 0$. Then define μ by

$$\mu(A) = \sum_{i=1}^n \mu(K_i).$$

It remains to show that μ is well defined and countably additive on $r(\mathcal{R})$. \square

As a consequence we obtain a more general result on uniqueness of the extension than the one in Theorem 2.8.3.

Lemma 2.8.5. *Let X be a nonempty set and let \mathcal{R} be a semi-ring of subsets of X such that X can be written as a countable union of elements of \mathcal{R} . Let μ be a countably additive set function on \mathcal{R} that is finite on \mathcal{R} . Then any measure ν on $\sigma(\mathcal{R})$ that agrees with μ on \mathcal{R} must agree with μ on $\sigma(\mathcal{R})$.*

Proof. To apply the monotone class theorem as in the proof of Theorem 2.8.3 we need \mathcal{R} to be an algebra. As \mathcal{R} is not necessarily an algebra our argument consists of techniques to reduce it to that case.

First we note that by Proposition 2.8.4 ν defined on \mathcal{R} has a unique extension to the ring generated by \mathcal{R} , so ν must agree with μ on $r(\mathcal{R})$. From the assumption on \mathcal{R} there exists a sequence of sets $X_n \in \mathcal{R}$, $\mu(X_n) < \infty$, such that $X_1 \subset X_2 \subset \cdots \subset X_n \subset \cdots$ and for any $A \in \sigma(\mathcal{R})$, $\nu(A) = \lim_{n \rightarrow \infty} \nu(A \cap X_n)$. Then, for each $n > 0$, ν is unique on the algebra in X_n generated by \mathcal{R} . Then the monotone class theorem argument applies. \square

The following observation is elementary but useful.

Lemma 2.8.6. *Let X be a nonempty set and let \mathcal{R} be a semi-ring of subsets of X such that X can be written as a countable union of elements of \mathcal{R} . Let μ be a countably additive set function on \mathcal{R} that is finite on elements of \mathcal{R} . Then \mathcal{R} is a sufficient semi-ring for the extension of μ as in Theorem 2.8.2.*

Before discussing some applications we prove a lemma that has some useful conditions to show that a finitely additive set function on a ring is countably additive.

Lemma 2.8.7. *Let \mathcal{R} be a ring of subsets of X and let μ be a finitely additive measure on \mathcal{R} .*

- (1) *If for all $A_i, A \in \mathcal{R}$ with $A_1 \subset A_2 \subset \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$ we have $\lim_{i \rightarrow \infty} \mu(A_i) = \mu(A)$, then μ is countably additive.*

(2) If μ is finite on \mathcal{R} and for all $A_i, A \in \mathcal{R}$ with $A_1 \supset A_2 \supset \cdots$ and $\bigcap_{i=1}^{\infty} A_i = \emptyset$ we have $\lim_{i \rightarrow \infty} \mu(A_i) = 0$, then μ is countably additive.

Proof. For part (1), if $\{B_i\}$ is any disjoint sequence of elements of \mathcal{R} with $A = \bigsqcup_{i=1}^{\infty} B_i \in \mathcal{R}$, then if $A_i = \bigsqcup_{i=1}^n B_i$ both A_i and A satisfy the hypotheses of the lemma and

$$\mu\left(\bigsqcup_{i=1}^{\infty} B_i\right) = \lim_{n \rightarrow \infty} \mu\left(\bigsqcup_{i=1}^n B_i\right) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu(B_i) = \sum_{i=1}^{\infty} \mu(B_i).$$

Finally, we show that, under the hypothesis of part (2), the hypothesis of part (1) holds. So let $\{A_i\}$ be an increasing sequence of sets in \mathcal{R} and $A = \bigcup_{i=1}^{\infty} A_i$. Set $B_i = A \setminus A_i$. Then $\{B_i\}$ is a decreasing sequence of sets and $\bigcap_{i=1}^{\infty} B_i = \emptyset$. Then $\mu(B_i) \rightarrow 0$, or $\mu(A \setminus A_i) \rightarrow 0$. As $\mu(A) < \infty$, this implies $\mu(A_i) \rightarrow \mu(A)$. \square

Application: Another Construction of Lebesgue measure on \mathbb{R}^d .

The reader is asked to prove the following lemma. (A similar lemma is shown in Lemma 4.9.1.)

Lemma 2.8.8. *The collection \mathcal{R}_d of all d -rectangles is a semi-ring.*

First one shows that the set function $|\cdot|_{(d)}$ defined earlier is finitely additive on the semi-ring of rectangles. The proof for $d = 1$ is similar to the proof of Lemma 2.1.2. The proof for $d > 1$ is reduced to the 1-dimensional case by the appropriate decomposition of the d -rectangle. The reader is asked to do this in the exercises (for the details in the 2-dimensional case refer to [68, Section 3.4]).

The next step is to use a compactness argument to show that $|\cdot|_{(d)}$ is countably additive on the semi-ring of rectangles. A proof of this is in [68, Section 3.4]. Then by Theorem 2.8.2 $|\cdot|_{(d)}$ extends to a unique measure on the Borel sets of \mathbb{R}^d .

Exercises

- (1) Let $X = \mathbb{N}$ and let \mathcal{C} consist of all subsets of \mathbb{N} . Define a set function on \mathcal{C} by $\mu(C) = \sum_{i \in C} \frac{1}{2^i}$ when C is a finite set and $\mu(C) = \infty$ when C is an infinite set. Show that μ is a finitely additive set function that is not countably additive.

- (2) Let μ be a countably additive set function on a semi-ring \mathcal{R} . Show that if $A, K_i \in \mathcal{R}$ and $A \subset \bigcup_i K_i$, then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(K_i)$.
- (3) In Lemma 2.8.7, part (2), give a direct proof that μ is countably additive without reducing it to part (1).
- (4) Show that $r(\mathcal{R})$, the ring generated by a semi-ring \mathcal{R} , consists of all finite unions of elements of \mathcal{R} .
- (5) Let X be a nonempty set and let \mathcal{R} be a semi-ring on X . Let μ be a set function on \mathcal{R} . Show that μ is countably additive if and only if it is additive and countably subadditive.
- (6) Show that if \mathcal{R} is a ring on a set X , then the collection $\mathcal{R} \cup \{X \setminus A : A \in \mathcal{R}\}$ is an algebra on X .
- (7) Complete the details in the proof of Lemma 2.8.5.
- * (8) Show that the d -volume $|\cdot|_d$ is finitely additive on the semi-ring of rectangles \mathcal{R}_d .
- * (9) Show that the Borel sets in \mathbb{R} are $\lambda_{(1)}^*$ -measurable where $\lambda_{(1)}^*$ is defined as in (2.7). Generalize this to $d > 1$.

Project A. We will extend the notion of outer measure on the line. A similar extension could be done on the plane. For any real numbers $0 < t \leq 1$ and $\delta > 0$ and any set A define

$$\mathcal{H}_\delta^t(A) = \inf \left\{ \sum_j |I_j|^t : A \subset \bigcup_j I_j \text{ and } |I_j| < \delta \right\}.$$

The t -dimensional Hausdorff measure of a set A is defined by

$$\mathcal{H}^t(A) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^t(A).$$

Observe that if $\delta_1 > \delta_2$, then $\mathcal{H}_{\delta_1}^t(A) < \mathcal{H}_{\delta_2}^t(A)$. Deduce that the limit above exists though it may be infinite. Show that $\mathcal{H}^t(A)$ satisfies the same properties as we showed Lebesgue outer measure satisfied.

The Hausdorff dimension of a set A is defined by

$$\dim_H(A) = \inf \{t \geq 0 : \mathcal{H}^t(A) = 0\}.$$

Show that $d = \dim_H(A)$ is such $\mathcal{H}^t(A) = \infty$ for $0 \leq t < d$ and $\mathcal{H}^t(A) = 0$ for $t > d$. Compute the Hausdorff dimension for some

examples. In particular show that the Cantor middle-thirds set has Hausdorff dimension $t = \frac{\log 2}{\log 3}$ and its $\frac{\log 2}{\log 3}$ -Hausdorff measure is 1.

Open Question A. This is an open question due to Erdős. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be the affine map defined by $f(x) = ax + b$ for any $a, b \in \mathbb{R}$. Two sets A and B in \mathbb{R} are said to be similar if $f(A) = B$ for some map f . Let A be a countable set in \mathbb{R} such that 0 is its only accumulation point. Is it the case that there exists a set $E \subset \mathbb{R}$ with $\lambda(E) > 0$, such that no subset of B is similar to A ? This question remains open but special cases are known. a) Search the literature to find out the special solutions to this question by Eigen and Falconer. b) Search the literature to find out what is currently known about this question. c) Write a paper describing in detail the partial solutions of Eigen and/or Falconer. d) Write a paper describing the state of the art on this question. Has this question been answered for the sequence $\{1/2^n : n > 0\}$? e) Solve the problem.

Chapter 3

Recurrence and Ergodicity

This chapter introduces two basic notions for dynamical systems: recurrence and ergodicity. As we saw in Chapter 1, at its most basic level, an abstract dynamical system with discrete time consists of a set X and a map or transformation T defined on the set X and with values in X . We think of X as the set of all possible *states* of the system and of T as the law of time evolution of the system. If the state of the system is x_0 at a certain moment in time, after one unit of time the state of the system will be $T(x_0)$, after two units of time it will be $T(T(x_0))$ (which we denote by $T^2(x_0)$), etc. We are first interested in studying what happens to states and to sets of states as the system evolves through time.

We shall study dynamical properties from a measurable or probabilistic point of view and thus impose a measurable structure on the set X and the map T . The basic structure on X is that of a measure space. While one can define many of the notions and prove some of the theorems in the setting of a general measure space (finite or σ -finite), for all the examples of interest and for many of the theorems with richer structure, one needs to assume some additional structure on X , such as that of a canonical Lebesgue measure space. We require T to be a measurable transformation, which we further specify

to be measure-preserving. We start, however, in Section 3.1, with an example that can be used to informally introduce some of the main ideas of this chapter and of Chapter 6.

Topological dynamics has been intimately connected with ergodic theory since the origins of both subjects, and we shall present some concepts from topological dynamics, such as minimality and topological transitivity.

3.1. An Example: The Baker's Transformation

We shall use the baker's transformation to introduce some important concepts in ergodic theory. These concepts are treated in this section in an informal manner and are studied in more detail in later sections.

When kneading dough, a traditional baker uses the following process: stretch the dough, then fold it and perform a quarter turn before repeating the stretch, fold and quarter turn process. The quarter turn is important when mixing a 3-dimensional piece of dough, but can be omitted in the 2-dimensional model. The 2-dimensional case already exhibits all the dynamical behaviors we are interested in but, before studying that, let us consider the 3-dimensional process without the quarter turn applied to the cube of Figure 3.1.

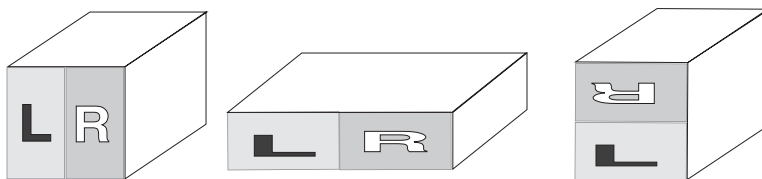


Figure 3.1. Kneading 3-d dough

Consider a right-handed 3-d coordinate system oriented so that the shaded face (marked L, R) of the cube in Figure 3.1 on the y, z -plane (the plane $x = 0$), with the bottom left corner of the part marked L at $(0, 0, 0)$, the bottom right corner of the part marked R at the point $(0, 1, 0)$, the top right corner of the part marked R at the point $(0, 1, 1)$, the top left corner of the part marked L at $(0, 0, 1)$ and the back corner of the cube that is not showing at the point $(-1, 0, 0)$,

so most of the cube, except for the shaded face, is on the negative x -axis. As the reader may verify, under this modified process each vertical cross section (i.e., a point in the plane $x = c$ ($0 \leq c \leq 1$)) remains invariant. So for example, points on the shaded face ($x = 0$) remain on the shaded face after an application of one iteration of the process; in fact, a point on the plane $x = b$ never reaches the plane $x = c$ for $b \neq c$. This clearly would not be a desirable process for mixing dough, as points on one half of the cube (points with $0 \geq x > -1/2$) never reach the other half (points with $-1/2 \geq x \geq -1$). However, as we shall see in Section 6.6 and explain informally in this section, there is mixing on each cross section of the form $x = a, 0 \geq a \geq 1$. As each cross section of the form $x = a, 0 \geq a \geq 1$, exhibits all the dynamical properties we wish to study, we shall confine our analysis to the 2-dimensional case from now on.

The modified 3-dimensional example demonstrates an important concept: each vertical cross section is mapped to itself, and so there are positive volume subsets of the cube that are mapped to themselves. A subset A of the cube is said to be *invariant* if the iterate (under the process) of every point (x, y, z) in A remains in the set A . Evidently, cross sections with $x = c$ are invariant, and so in particular half the cube (the subset constrained by $0 \leq x \leq 1/2$) is invariant. An invariant set A gives rise to a new dynamical system as one can consider the transformation restricted to the invariant set as a dynamical system in its own right. It can be shown, however, that the original process, including the quarter turn, on the cube does not have any invariant sets of positive volume.

To describe the 2-dimensional process, we look more carefully at the action of the modified 3-dimensional process when applied to 2-dimensional cross sections of the form $x = a, 0 \geq a \geq 1$, which are squares of side length 1. We concentrate on the cross section at $x = 0$ and have shaded each half of it with different intensity to better visualize the process. One full cycle of the kneading process now consists of a stretch and then a fold; stretching produces a figure as in the right; then there is a folding that occurs to return to the original shape.

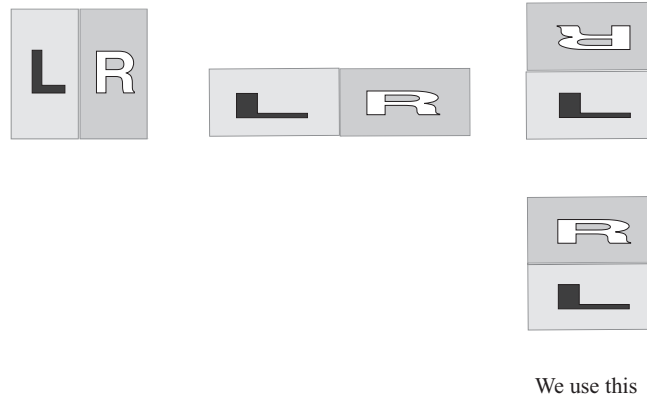


Figure 3.2. Kneading 2-d dough

We note that, as shown in Figure 3.1, the right rectangle ends up upside-down. For the sake of mathematical convenience, instead of putting the top piece upside-down, we will put it rightside-up. (From the point of view of the dynamical ideas we discuss, this change gives an equivalent process.) We are ready to define the 2-dimensional baker's transformation in detail. Start with a unit square. Cut the square down the middle to obtain two equal pieces (subrectangles). Then squeeze and stretch the left piece and do the same with the right piece; finally put the right piece on top of the left to bring us back to a square. This concludes one iteration of the process (Figure 3.2) and defines a transformation T of the unit square. It is easy to see that if a point is in the left subrectangle, then its horizontal distance is doubled and its vertical distance is halved, and points in the right subrectangle undergo a similar transformation, so T is given by the following formula:

$$T(x, y) = \begin{cases} (2x, \frac{y}{2}), & \text{if } 0 \leq x < 1/2; \\ (2x - 1, \frac{y+1}{2}), & \text{if } 1/2 \leq x \leq 1. \end{cases}$$

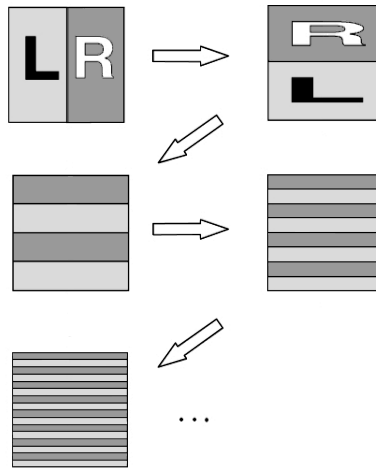


Figure 3.3. The baker's transformation

The first measurable dynamical property that T satisfies is that it preserves the area of subsets of the square. (Such a transformation is called *measure-preserving*.) Note that the rectangle L of area $1/2$ is sent to another rectangle $T(L)$ also of area $1/2$. The same is true for R and, furthermore, one can easily calculate that any rectangle whose sides are bounded by the *dyadic rationals* (i.e., numbers of the form $\frac{k}{2^n}$ for k and n integers) is sent to another rectangle of the same area. One can show that this property holds for arbitrary rectangles by approximating them with arbitrary dyadic rectangles. It is reasonable to ask what happens to more general subsets of the square. To this end, Chapter 2 develops the theory of Lebesgue measure, which generalizes the notion of area for the case of subsets of the plane. Our initial treatment of Lebesgue measure is restricted to subsets of the real line, but the same ideas generalize to subsets of \mathbb{R}^d . We show in Section 3.4 that to establish the measure-preserving property it suffices to verify it on a sufficiently large collection of sets, such as the collection of all rectangles with dyadic endpoints. We note here that in the definition of the measure-preserving property, as we shall

see later, the condition is that the measure of the inverse image of a set is the same as the measure of the set; when the process or transformation is invertible, as is the case here, this is equivalent to the condition that the *forward* image of a set is the same as the measure of the set.

Studying what happens to specific sets when we iterate the baker's transformation helps understand the notion of mixing. Consider the set L (Figure 3.3). It is clear that $T(L)$ intersects L and R each in a square of area $1/4$; in other words, if we let $\lambda(X)$ denote the area (or measure) of a set X of the plane, we have that

$$\lambda(T(L) \cap A) = \lambda(L \cap A),$$

where A is L or R . The reader is asked in the exercises to verify that for any sets A and B that are dyadic rectangles,

$$\lambda(T^n(A) \cap B) = \lambda(A)\lambda(B),$$

for all integers $n \geq 0$. For more general sets, namely for any measurable sets A and B , it can be shown by approximating them by unions of dyadic rectangles and using the techniques of Chapter 6 that

$$(3.1) \quad \lim_{n \rightarrow \infty} \lambda(T^n(A) \cap B) = \lambda(A)\lambda(B),$$

which, when $\lambda(B) \neq 0$, can be written as

$$(3.2) \quad \lim_{n \rightarrow \infty} \frac{\lambda(T^n(A) \cap B)}{\lambda(B)} = \lambda(A).$$

This is the definition of *mixing* for a measure-preserving transformation. We interpret equation (3.2) as saying that the relative proportion in which the iterates of a given set A intersect a region of space B approximates the measure of A (Figure 3.4). For example, if in the kneading dough process one were to drop in some raisins occupying 5% of the area, and if we fix our gaze on a particular part of the space, then after some iterations one would expect to see approximately 5% of this region occupied by raisins.

Another important concept arises when we relax the notion of convergence in equation (3.1). It may be that convergence in the sense of (3.1) does not occur, but one may still have convergence in

the average, namely it may happen that for all measurable sets A and B it is the case that

$$(3.3) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} \lambda(T^i(A) \cap B) = \lambda(A)\lambda(B).$$

This is a weaker notion, and if (3.3) holds (for all measurable sets) we say that the transformation T is *ergodic*. We shall see that ergodicity is a weaker notion than mixing. It is also interesting to note, though we do not cover this, that by putting a topology on the set of measure-preserving transformations defined on a space such as the unit square, it can be shown that generically (i.e., for “a large set” in the sense of the topology) one will pick a transformation that is ergodic but not mixing. Another characterization of ergodicity can be seen by studying invariant sets: if there were to exist an invariant set A such that both A and its complement A^c had positive measure, then T would not be ergodic, because then $\lambda(T^n(A) \cap A^c) = 0$, contradicting (3.3) when $B = A^c$. The converse of this is also true but nontrivial and is a consequence of the ergodic theorem, which we prove in Chapter 5.

Returning to the modified 3-dimensional baker's transformation, as we saw that it admits an invariant set of positive measure whose complement is also of positive measure, it follows that the transformation is not ergodic. This example also shows an important property of such transformations. We have seen that the cube on which the modified 3-dimensional transformation is defined can be decomposed into cross sections of the form $x = c$ that are invariant, and the process is ergodic (in fact, mixing) with respect to 2-dimensional measure when restricted to each 2-dimensional slice. This is called the ergodic decomposition of the transformation and it is a special case of an important theorem called the ergodic decomposition theorem. In general terms, this theorem states that any measure-preserving transformation has a decomposition into ergodic components. Because of this theorem, when proving facts about transformations it is often possible to simply prove the result for the case of ergodic transformations.

This is a good place to mention a very useful property that lies properly between ergodicity and mixing. It may happen that the limit in (3.1) holds along a subsequence, that is, for each pair of measurable

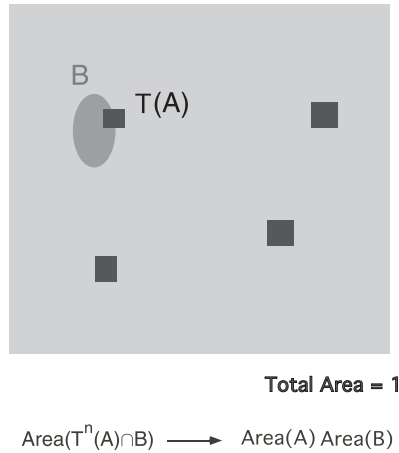


Figure 3.4. Mixing

sets A and B there may exist an increasing sequence n_i so that

$$(3.4) \quad \lim_{i \rightarrow \infty} \lambda(T^{n_i}(A) \cap B) = \lambda(A)\lambda(B).$$

In this case we shall say that the transformation T is *weakly mixing*. Evidently, mixing implies weak mixing, and it is not hard to see that weak mixing implies ergodicity. In Chapter 6 we show that there exists a sequence n_i that “works” for all pairs of measurable sets A and B . Surprisingly, one can choose the sequence n_i to be of density 1 in the integers. Examples of transformations that are weakly mixing but not mixing are not immediately obvious and we shall construct some. Ergodic transformations that are not weakly mixing are easier to come by.

Exercises

- (1) Show that in the case of the 2-dimensional baker’s transformation T , for each dyadic rectangle A in the unit square, A and $T(A)$ have the same area.
- (2) Verify that the mixing condition implies the ergodic condition.

- (3) State and prove the analogue of Exercise 1 for the case of the 3-dimensional baker's transformation.

3.2. Rotation Transformations

This section introduces one of the simplest examples of a transformation that leaves the Lebesgue measure of the unit interval invariant; such a transformation is said to be a *measure-preserving* transformation. A **transformation** is a function for which the domain and range are the same; in this case the image of a point is in the domain of the transformation. The simplest transformation on any set X is the identity transformation \mathcal{I} defined by $\mathcal{I}(x) = x$ for all $x \in X$. If $T : X \rightarrow X$ is a transformation on a set X , as $T(x) \in X$ for all $x \in X$, the n^{th} iterate of x , denoted $T^n(x)$, is defined by

$$\begin{aligned} T^0 &= \mathcal{I}, \\ T^{n+1} &= T \circ T^n \text{ for } n \geq 0. \end{aligned}$$

An **invertible transformation** is a transformation that is one-to-one and onto. In this case T^{-1} is also a transformation.

The rotation transformations we consider in this section are defined on the half-open unit interval $[0, 1)$. First we assign to each x in \mathbb{R} a unique number in $[0, 1)$, denoted $(x \bmod 1)$ and defined by

$$(x \bmod 1) = x - \lfloor x \rfloor,$$

where $\lfloor x \rfloor$ is the largest integer $\leq x$. (So, $(x \bmod 1) = x$ if $x \in [0, 1)$ and, for example, $(\pi \bmod 1) = 0.1415\dots$) Using this one can define an equivalence relation on \mathbb{R} : two numbers $x, y \in \mathbb{R}$ are said to be **equivalent mod 1** and written $x \equiv y \pmod{1}$ if $(x - y \bmod 1) = 0$.

For any number $\alpha \in \mathbb{R}$, define the **rotation by α** to be the transformation

$$R_\alpha : [0, 1) \rightarrow [0, 1)$$

given by

$$R_\alpha(x) = (x + \alpha \bmod 1).$$

When α is fixed and evident from the context we shall write R for R_α . Clearly, R_α is an invertible transformation.

Question. Show that for any $\alpha \in \mathbb{R}$ there exists $\alpha' \in [0, 1)$ such that $R_\alpha = R_{\alpha'}$.

From now on we assume that $\alpha \in [0, 1)$.

Example. If $\alpha = \frac{3}{4}$, then $R_\alpha(0) = \frac{3}{4}$, $R_\alpha^2(0) = R_\alpha(\frac{3}{4}) = (\frac{3}{4} + \frac{3}{4} \bmod 1) = \frac{1}{2}$, $R_\alpha^3(0) = (\frac{1}{2} + \frac{3}{4} \bmod 1) = \frac{1}{4}$, $R_\alpha^4(0) = R_\alpha(\frac{1}{4}) = 0$.

This example motivates one of the first notions in dynamics. If $T : X \rightarrow X$ is a transformation, then we call the set $\{T^n(x)\}_{n \geq 0}$ the **positive orbit** of x . When T is invertible we usually consider the **full orbit** of a point x , namely the set $\{T^n(x)\}_{n=-\infty}^{\infty}$.

Question. Show that if α is a rational number, then the orbit of every point x in $[0, 1)$ under R_α consists of a finite number of points.

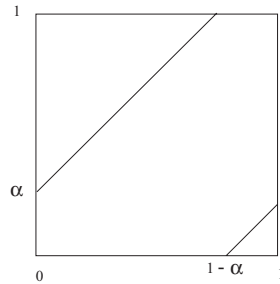
We say that a point $x \in X$ is a **periodic point** under a transformation $T : X \rightarrow X$ if $T^n(x) = x$ for some integer $n > 0$. The integer n is called a **period** of x ; the **least period** is the smallest such integer n . A transformation is said to be a **periodic transformation** if every point is a periodic point for T ; it is **strictly periodic** if every point has the same period. For example, $R_{3/4}$ is a strictly periodic transformation of least period 4.

Figure 3.5 shows the graph of R_α . The figure also suggests a description of R_α as an “interval exchange.” Note that there are two partitions of $[0, 1)$, one consisting of the subintervals $[0, 1 - \alpha)$ and $[1 - \alpha, 1)$, and the other of the subintervals $[0, \alpha)$ and $[\alpha, 1)$. The transformation R_α sends an interval in the first partition to an interval of the same length in the second partition.

Observe that

$$R_\alpha(0) = \alpha, R_\alpha(1 - \alpha) = 0.$$

Also, when $x < 1 - \alpha$, then $x + \alpha < 1$, so $R_\alpha(x) = (x + \alpha \bmod 1) = x + \alpha$. Therefore, R_α on $[0, 1 - \alpha)$ is just the translation that sends x to $x + \alpha$ and whose graph is the first straight line of Figure 3.5. Now if $1 - \alpha < x < 1$, then $(x + \alpha \bmod 1) = x + \alpha - 1$ and the graph of R_α is the second line of Figure 3.5. (Note that in the graph all

Figure 3.5. Rotation by R_α

intervals are open at 1.) From this it is clear that R_α is an invertible transformation on $([0, 1], \mathfrak{L}, \lambda)$.

Given a transformation $T : X \rightarrow X$ and a set $A \subset X$, we shall be interested in the set of points x such that $T(x) \in A$. Recall that the **inverse image** (or **pre-image**) of A is defined to be the set

$$T^{-1}(A) = \{x : T(x) \in A\}.$$

So, $T(x) \in A$ if and only if $x \in T^{-1}(A)$. In addition to considering the iterates of points $x, T(x), T^2(x), \dots$ we consider the pre-images $T^{-n}(A), n \geq 0$. Note that $T^n(x) \in A$ if and only if $x \in T^{-n}(A)$. When T is invertible we shall also consider $T^n(A)$ for $n \geq 0$.

We now introduce the following definitions, which will be discussed in more detail in Section 3.4.

If (X, \mathcal{S}, μ) is a measure space, a transformation $T : X \rightarrow X$ is said to be a **measurable transformation** if the set $T^{-1}(A)$ is in $\mathcal{S}(X)$ for all $A \in \mathcal{S}(X)$. An **invertible measurable transformation** is an invertible transformation T such that T and T^{-1} are measurable. A transformation T is **measure-preserving** if it is measurable and

$$\mu(T^{-1}(A)) = \mu(A)$$

for all sets $A \in \mathcal{S}(X)$. In this case we say that μ is an **invariant measure** for T . Thus, a measure-preserving transformation must be

measurable, but sometimes for emphasis we may describe a transformation as being measurable and measure-preserving. (An **invertible measure-preserving** transformation is an invertible measurable transformation that is measure-preserving.)

Question. Show that for any measure space (X, \mathcal{S}, μ) the identity transformation $\mathfrak{J} : X \rightarrow X$, defined by $\mathfrak{J}(x) = x$, is an invertible measure-preserving transformation.

Question. Let T be an invertible measurable transformation. Show that T is measure-preserving if and only if T^{-1} is measure-preserving.

Theorem 3.2.1. *The transformation R_α is an invertible measure-preserving transformation of $([0, 1], \mathfrak{L}, \lambda)$.*

Proof. Recall that we may assume $\alpha \in [0, 1)$. Note that the inverse of R_α is $R_\alpha^{-1} = R_{-\alpha}$, another rotation. Put $A_0 = A \cap [0, \alpha)$ and $A_1 = A \cap [\alpha, 1)$. Then note that $R_{-\alpha}(A_0)$ is precisely the translation by $-\alpha$ of the set A_0 , which by Exercise 2.3.1 is measurable and of the same measure as A_0 . A similar argument applies to A_1 and, since $A = A_0 \sqcup A_1$, this completes the proof. \square

As we shall see later, the proof of the measure-preserving property for a given transformation will usually be obtained as a consequence of Theorem 3.4.1, by which it suffices to show that $R_\alpha^{-1}(A)$ is measurable and that $\lambda(R_\alpha^{-1}(A)) = \lambda(A)$ for all measurable sets A in some sufficient semi-ring. For example, in the case of Theorem 3.2.1, the sufficient semi-ring can be the collection of intervals contained in $[0, 1)$.

Our next theorem, Kronecker's theorem for irrational rotations, has several applications. Before we discuss its proof, we need to introduce some notation. We represent rotations as being on the unit interval, rather than on the unit circle, but we identify 0 with 1. Under this identification, the number 0.1 is closer to 0.9 than to 0.4. We now define a metric, or distance, that reflects the fact that we are considering numbers modulo 1. For $x, y \in [0, 1)$ define $d(x, y)$ by

$$d(x, y) = \min\{|x - y|, 1 - |x - y|\}.$$

Note that if $|x - y| \leq 1/2$, then $d(x, y) = |x - y|$.

The proof of the following proposition is left to the reader.

Proposition 3.2.2. *The function d is a metric on $[0, 1]$, and is such that if a sequence converges in the Euclidean metric, then it converges in the d metric. Furthermore, d is invariant for rotations, i.e., for any rotation R_α and any $x, y \in [0, 1]$, $d(R_\alpha(x), R_\alpha(y)) = d(x, y)$.*

Theorem 3.2.3 (Kronecker). *If α is irrational, then for every $x \in [0, 1]$ the sequence $\{R_\alpha^n(x)\}_{n \geq 0}$ is dense in $[0, 1]$.*

Proof. We first show that when α is irrational, all the points $R_\alpha^n(x)$ are distinct for different integers n . Let R denote R_α . Now if $R^n(x) = R^m(x)$, for some integers m, n , then $x + n\alpha \equiv x + m\alpha \pmod{1}$, so $(n - m)\alpha \equiv 0 \pmod{1}$, which implies that $(n - m)\alpha$ is an integer. As α is irrational, this happens only when $m = n$. Therefore all the points in the orbit of x are distinct.

By the Bolzano–Weierstrass theorem we know that the sequence $\{R^n(x)\}_{n \geq 0}$ has a convergent subsequence in $[0, 1]$. This implies that there are points in the orbit that are arbitrarily close to each other: given any $1/2 > \varepsilon > 0$ there exist nonnegative integers $p > q$ such that $0 < |R^p(x) - R^q(x)| < \varepsilon$. As $\varepsilon < 1/2$, $d(R^p(x), R^q(x)) < \varepsilon$. By Proposition 3.2.2, $d(R^k(y_1), R^k(y_2)) = d(y_1, y_2)$ for all $y_1, y_2 \in [0, 1]$ and all integers k . So, $d(R^{p-q}(x), x) < \varepsilon$. Let $r = p - q$ and $\delta = d(R^{p-q}(x), x) > 0$. We claim that consecutive terms in the orbit $\{R^{\ell r}(x)\}_{\ell \geq 0}$ are δ -apart of each other. In fact, for $\ell \geq 0$,

$$d(R^{(\ell+1)r}(x), R^{\ell r}(x)) = d(R^r(x), x) = \delta < \varepsilon.$$

This shows that the distinct points $\{R^{\ell r}(x)\}_{\ell=0}^\infty$ subdivide $[0, 1]$ into subintervals of length $< \varepsilon$. Therefore, $\{R^n(x)\}_{n \geq 0}$ is dense in $[0, 1]$. \square

The notion of continuity of real functions, with which we assume the reader is familiar, generalizes in a natural way to metric spaces. Let (X, d) and (Y, q) be two metric spaces. A map $\phi : X \rightarrow Y$ is said to be **continuous at a point** $x \in X$ if for all $\varepsilon > 0$ there exists a number $\delta > 0$ such that $q(\phi(x), \phi(y)) < \varepsilon$ whenever $y \in X$ and $d(x, y) < \delta$. We say that f is **continuous** on X if f is continuous at x for every point x in X . The reader is asked to show

that R_α is continuous with respect to the d metric of Proposition 3.2.2 (Exercise 6).

Theorem 3.2.3 illustrates an important concept. If (X, d) is a metric space, a map $T : X \rightarrow X$, usually continuous, is defined to be **minimal** if the positive orbit $\{T^n(x)\}_{n \geq 0}$ is dense in X for all $x \in X$. Kronecker's theorem states that irrational rotations are minimal. Minimality is not invariant under (measure-preserving) isomorphisms (defined in Section 3.10).

Application 1. We apply Theorem 3.2.3 to an interesting question known as Gelfand's question. This question is concerned with the first digits of powers of 2. Here is a list of the first 20 powers of 2:

2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192,
16384, 32768, 65536, 131072, 262144, 524288, 1048576

The sequence of first digits of the first 40 powers of 2 is:

2, 4, 8, 1, 3, 6, 1, 2, 5, 1,
2, 4, 8, 1, 3, 6, 1, 2, 5, 1,
2, 4, 8, 1, 3, 6, 1, 2, 5, 1,
2, 4, 8, 1, 3, 6, 1, 2, 5, 1.

Do we ever see a 7, a 9? Gelfand's question asks: how often do we see a power of 2 that starts with a 7, and with what frequency? We show here that there are infinitely many integers n such that 2^n starts with a 7. Surprisingly, they have a well-defined frequency. The existence of this frequency follows from the uniform distribution of multiples of an irrational number modulo 1. While this fact can be given an independent proof, we will obtain it as a consequence of the ergodic theorem, which we study in Chapter 5. We shall see later in Section 5.3, that the digit $d, 0 < d < 10$, occurs as a first digit in powers of 2 with frequency $\log_{10}(d+1)/d$. So for example the digit 1 appears with frequency about 0.3. Benford's Law is a statement asserting that the digit 1 occurs as a first digit in many naturally occurring tables, such as street addresses, stock prices, electric bills, with frequency about 1/3 (rather than about 1/9 as might be expected), and that the frequency of the other digits continues to be distributed in this logarithmic form. This was first observed by S. Newcomb in 1881

regarding first digits in logarithm tables, but only justified recently in 1996 by T. Hill.

We start with the following simple observation. As the integer 721, say, starts with a 7 there is some integer $k > 0$ so that

$$7 \times 10^k \leq 721 < 8 \times 10^k.$$

To generalize this, let $d \in \{1, \dots, 9\}$ (a similar analysis can be done for any integer $d > 0$, but the details are left to the reader). If the decimal representation of the integer 2^n starts with d , then for some integer $k \geq 0$,

$$d \times 10^k \leq 2^n < (d + 1) \times 10^k.$$

Thus

$$\log_{10}(d \times 10^k) \leq \log_{10} 2^n < \log_{10}((d + 1) \times 10^k).$$

In other words,

$$\begin{aligned} \log_{10} d &\leq n \log_{10} 2 - k < \log_{10}(d + 1), \text{ or} \\ \log_{10} d &\leq n \log_{10} 2 \pmod{1} < \log_{10}(d + 1). \end{aligned}$$

But this is the same as saying that, letting $\alpha = \log_{10} 2$,

$$R_\alpha^n(0) \in [\log_{10} d, \log_{10}(d + 1)).$$

Since $\alpha = \log_{10} 2$ is irrational, by Theorem 3.2.3 there are infinitely many integers n such that $R_\alpha^n(0) \in [\log_{10} d, \log_{10}(d + 1))$. Thus there are infinitely many powers of 2 that start with a 7.

Application 2. (A Nonmeasurable Set.) The following construction of a nonmeasurable set is based on the Axiom of Choice. We first recall the statement of this axiom: Given any collection of nonempty sets A_α indexed by some nonempty set Γ , there exists a function F , called a choice function, whose domain is Γ and whose range is $\bigcup_{\alpha \in \Gamma} A_\alpha$, such that $F(\alpha) \in A_\alpha$, for each $\alpha \in \Gamma$. We think of F as choosing an element of A_α for each $\alpha \in \Gamma$. If the set Γ were a finite set, using mathematical induction on the integers, it is not hard to give a proof for this axiom. However, P.J. Cohen showed in 1963 that the Axiom of Choice is independent of the standard axioms of set theory (the Zermelo-Fraenkel axioms), i.e., neither it nor its negation can be deduced from these axioms. Furthermore, in 1970 Solovay showed that it cannot be shown that there exist nonmeasurable sets of reals

with the Zermelo-Fraenkel axioms of set theory without the Axiom of Choice, on the assumption that the existence of *inaccessible cardinals* is consistent with the Zermelo-Fraenkel axioms. The Axiom of Choice is a reasonable axiom that is assumed by most mathematicians.

We use irrational rotations to show the existence of a non-measurable set. Let R be any rotation by an irrational number. Then for any $x \in [0, 1)$ the full orbit $\Gamma_x = \{R^n(x)\}_{n=-\infty}^{\infty}$ consists of distinct points. We claim that the collection of orbits forms a partition of $[0, 1)$, i.e., we claim that every point in $[0, 1)$ is in some orbit and that if any two orbits intersect in a nonempty set, then they must be equal. Indeed, for any $x \in [0, 1)$, $x \in \Gamma_x$, and if $z \in \Gamma_x \cap \Gamma_y$, for some $x, y \in [0, 1)$, then $z = R^n(x)$ and $z = R^m(y)$ for some integers m, n . So $x = R^{m-n}(y)$ which means that $\Gamma_x = \Gamma_y$. (Another way to obtain this partition is to define an equivalence relation on $[0, 1)$ by declaring that two points x, y are equivalent if $x \in \Gamma_y$. One verifies that this is an equivalence relation and that the equivalence classes are the orbits.)

As each orbit is nonempty, we can use the Axiom of Choice to construct a set E consisting of precisely one point from each orbit. It follows that for each integer n , $R^n(E)$ also consists of one point from each orbit, and this means that the collection $\{R^n(E)\}_{n=-\infty}^{\infty}$ forms a countable partition of $[0, 1)$ into disjoint sets.

We now show that assuming that the set E is measurable leads to a contradiction. So suppose that E is measurable. Since R is measure-preserving, $R^n(E)$ is measurable and $\lambda(R^n(E)) = \lambda(E)$ for all n . Since $\bigcup_{n=-\infty}^{\infty} R^n(E) = [0, 1)$, by Countable Additivity

$$1 = \sum_{n=-\infty}^{\infty} \lambda(R^n(E)) = \sum_{n=-\infty}^{\infty} \lambda(E).$$

But if $\lambda(E) = 0$, then $\sum_{n=-\infty}^{\infty} \lambda(E) = 0$ and if $\lambda(E) > 0$, then $\sum_{n=-\infty}^{\infty} \lambda(E) = \infty$. So in both cases we reach a contradiction. Therefore the set E is not measurable.

Exercises

- (1) Let d be any positive integer. Show that there are infinitely many positive integers n so that 2^n starts with d .

- (2) Show that there are infinitely many positive integers n so that 3^n starts with 1984.
- (3) Let $X_n = \{x_0, \dots, x_{n-1}\}$ and let μ be the counting measure on X_n , i.e., $\mu(\{x_i\}) = 1$ for $i = 0, \dots, n-1$. Define a transformation T on X_n by $T(x_i) = x_{i+1}$ if $i = 0, \dots, n-2$, and $T(x_{n-1}) = x_0$. Show that T is a measure-preserving invertible transformation on (X_n, μ) . T is called a rotation on n points.
- (4) Give another proof of Kronecker's Theorem by showing that if the orbit of a point $x \in [0, 1)$ is not dense, then one can choose an open interval in the complement of the orbit and then reach a contradiction.
- (5) Prove Proposition 3.2.2.
- (6) Show that R_α is continuous with respect to the metric d .
- (7) Show that, assuming the Axiom of Choice, every interval contains a nonmeasurable set. (In Exercise 3.11.1 you will show that every set of positive measure contains a non-measurable set.)
- (8) Is it the case that for every $\varepsilon > 0$ there exists a non-Lebesgue measurable set E with $\lambda^*(E) < \varepsilon$?
- * (9) A subset E of \mathbb{R} is called a **Bernstein set** if both E and E^c have a nonempty intersection with each uncountable closed set. Show that a Bernstein set, if it exists, is non-Lebesgue measurable. (For a construction of Bernstein sets see [56].)
- * (10) Are there infinitely many positive integers n so that both 2^n and 3^n start with 7?

3.3. The Doubling Map: A Bernoulli Noninvertible Transformation

The second example that we study is another transformation defined on $[0, 1)$; however this time it is not invertible, but two-to-one. Define the transformation T on $[0, 1)$ by

$$T(x) = 2x \pmod{1} = \begin{cases} 2x, & \text{if } 0 \leq x \leq 1/2; \\ 2x - 1, & \text{if } 1/2 < x < 1. \end{cases}$$

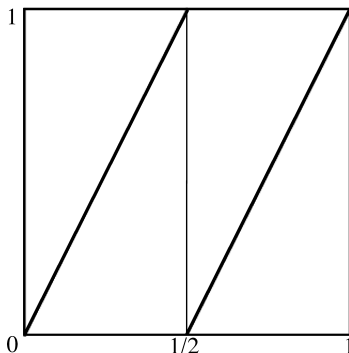


Figure 3.6. The doubling map

One can think of T as a one-dimensional version of the baker's transformation. We call T the **doubling map** transformation. Before studying its properties we investigate a new representation of the doubling map.

Let D consist of all the numbers in $[0, 1]$ of the form $\frac{k}{2^n}$, and let $X_0 = [0, 1] \setminus D$. The numbers in $[0, 1] \setminus D$ have a unique representation in binary form as

$$x = \sum_{i=1}^{\infty} \frac{a_i}{2^i}.$$

We call the sequence $(a_1 a_2 \dots)$ the **symbolic binary representation** of x . Then the doubling map for points $x \in [0, 1] \setminus D$ is given by

$$T(x) = \sum_{i=1}^{\infty} \frac{a_{i+1}}{2^i}.$$

That is, T is a *shift* of the representation of x , i.e., the point in $[0, 1)$ whose symbolic binary representation is $(a_1 a_2 a_3 \dots)$ is sent to the point with symbolic binary representation $(a_2 a_3 \dots)$. The following theorem shows that the doubling map is measure-preserving. The use of the inverse image of sets in the definition of the measure-preserving property is further clarified in Lemma 4.4.6.

Theorem 3.3.1. *The doubling map transformation T is a measure-preserving transformation on $([0, 1), \mathcal{L}, \lambda)$. Furthermore, the set of periodic points of T is dense in $[0, 1)$.*

Proof. We first show it is measure-preserving. Define the maps $S_1 : [0, 1) \rightarrow [0, \frac{1}{2})$ and $S_2 : [0, 1) \rightarrow [\frac{1}{2}, 1)$ by $S_1(y) = y/2$ and $S_2(y) = y/2 + 1/2$. We saw in Exercise 2.3.1 that for any measurable set A , the sets $\frac{1}{2}A$ and $\frac{1}{2}A + \frac{1}{2}$ are measurable and $\lambda(\frac{1}{2}A) = \frac{1}{2}\lambda(A) = \lambda(\frac{1}{2}A + \frac{1}{2})$. Next we observe that $T^{-1}(A) = S_1(A) \sqcup S_2(A)$. So $\lambda(T^{-1}(A)) = \lambda(S_1(A)) + \lambda(S_2(A)) = \lambda(\frac{1}{2}A) + \lambda(\frac{1}{2}A + \frac{1}{2}) = \frac{1}{2}\lambda(A) + \frac{1}{2}\lambda(A) = \lambda(A)$.

A second proof can be given using Theorem 3.4.1.

For the periodic points, one can verify that for each integer $p \geq 1$ the points of period p are those whose symbolic binary representation consists of the infinitely repeated string $a_1a_2 \dots a_p$ for any $a_1, \dots, a_p \in \{0, 1\}$. The set of these points is clearly dense in $[0, 1)$. \square

Application. From the representation of T as a shift of the binary numbers in $[0, 1)$, it follows that

$$T^i(x) \in [0, 1/2)$$

if and only if the i^{th} digit in the binary expansion of x is 0, and

$$T^i(x) \in [1/2, 1)$$

if and only if the i^{th} digit in the binary expansion of x is 1.

Using the important **characteristic function** or **indicator function** of a set A , denoted by \mathbb{I}_A and defined by

$$\mathbb{I}_A(x) = \begin{cases} 1, & \text{if } x \in A; \\ 0, & \text{if } x \notin A, \end{cases}$$

we see that the i^{th} digit in the binary expansion of x is 1 if and only if $T^i(x) \in [1/2, 1)$, and it is 0 if and only if $T^i(x) \in [0, 1/2)$.

Therefore, the frequency of appearances of 0 in the binary representation of x , if it exists, can be expressed as the limit

$$(3.5) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^n \mathbb{I}_{[0, 1/2)}(T^i(x)).$$

Limits of the above form will be studied in Chapter 5, and their existence, for all x outside a set of measure zero, will be a consequence of the Ergodic Theorem. A number x is said to be a **simply normal number** to the base 2 if the limit in (3.5) exists and equals $1/2$. A consequence of the ergodic theorem will be that almost all (i.e.,

outside a null set) numbers are *normal*, an extension of simply normal defined in Section 5.2. Normal numbers are discussed in more detail in Section 5.2.

Note that the only point of discontinuity of T is at $x = 1/2$. However, with respect to the metric d of Section 3.2, T is continuous. We end with another example that has similar properties but is continuous with respect to the Euclidean metric. Define the **tent map** $T : [0, 1] \rightarrow [0, 1]$ by

$$T(x) = \begin{cases} 2x, & \text{if } x \in [0, 1/2); \\ 2 - 2x, & \text{if } x \in [1/2, 1]. \end{cases}$$

Let (X, d) be a metric space and let $T : X \rightarrow X$ be a map, usually a continuous map. The map T is **topologically transitive** (sometimes called one-sided topologically transitive) if there is a point $x \in X$ such that its positive orbit $\{T^n(x) : n \geq 0\}$ is dense in X . For example, the tent map T is topologically transitive (Exercise 8). Also, any minimal map is topologically transitive.

We conclude with a very remarkable theorem that has proven very useful in establishing existence results in dynamics. We first need a few definitions. The notion of nowhere dense sets that we have seen earlier can be generalized to metric spaces. Let (X, d) be a metric space. A set A in X is said to be **nowhere dense** if its closure contains no nonempty open sets; we saw, for example, that the Cantor set in $[0, 1]$ is nowhere dense. We observe that A is nowhere dense if and only if for any nonempty open set G there is a nonempty open set H contained in $G \setminus A$. In fact, suppose A is nowhere dense and G is a nonempty open set. If $G \setminus A$ were to contain no open sets, then A would be dense in G , or the closure of A would contain G , a contradiction. Conversely, if for any nonempty open set G there is a nonempty open set $H \subset G \setminus A$, then A cannot be dense in any open set G , which means that its closure contains no nonempty open sets.

A set is said to be **meager** or of **first category** if it is a countable union of nowhere dense sets; for example, the set of rational numbers in \mathbb{R} , while not nowhere dense is a meager set. The concept of meager captures the notion of being topologically small.

The next theorem basically states that a complete metric space cannot be topologically small. In fact, it says that a nonempty open set in a complete metric space cannot be meager.

Lemma 3.3.2. *Let (X, d) be a metric space. Then the following are equivalent.*

- (1) *Every nonempty open set is not a meager set.*
- (2) *The intersection of any countable collection of dense open sets is dense.*

Proof. Suppose that a nonempty open set is not meager. Let $A = \bigcap_{i=1}^{\infty} G_i$, where each G_i is an open dense set. Then G_i^c is nowhere dense. If A were not dense there would exist a nonempty open set H contained in A^c . But $A^c = \bigcup_{i=1}^{\infty} G_i^c$, a meager set. This contradicts that it contains H .

For the second part suppose that H is an open set that is meager. Then $H = \bigcup_{i=1}^{\infty} E_i$, where the E_i are nowhere dense sets. Then their closures \bar{E}_i are also nowhere dense. Then $H \subset \bigcup_{i=1}^{\infty} \bar{E}_i$, so H^c contains $\bigcap_{i=1}^{\infty} \bar{E}_i^c$, a countable intersection of dense open sets, which cannot be dense as it does not intersect H , a contradiction. \square

Theorem 3.3.3 (Baire Category Theorem). *Let X be a complete metric space. Then the intersection of any countable collection of dense open sets in X is dense.*

Proof. Let $\{G_n\}$ be a countable collection of dense open sets in X . We show that if B is any open ball of positive radius, then it intersects $\bigcap_{n=1}^{\infty} G_n$. The idea is to construct a point in the intersection as the limit of a Cauchy sequence.

We start the construction of the sequence. For this we have to be careful to choose the sequence inside balls of decreasing radius ε_n to guarantee it is Cauchy—for example, it suffices to take $\varepsilon_n < \frac{1}{n}$. First note that as G_1 is dense, B has a nonempty intersection with G_1 . Choose a ball, of the form $B(x_1, \varepsilon_1)$, so that its closure is contained in $B \cap G_1$ and with $\varepsilon_1 < 1$. The ball $B(x_1, \varepsilon_1)$ must also intersect G_2 and we may similarly choose a ball $B(x_2, \varepsilon_2)$ so that its closure is contained in $B(x_1, \varepsilon_1) \cap G_2$ and with $\varepsilon_2 < 1/2$. In this way we can generate a sequence of balls $B(x_n, \varepsilon_n)$ so that $\bar{B}(x_n, \varepsilon_n) \subset B(x_{n-1}, \varepsilon_{n-1}) \cap G_n$

and $\varepsilon_n < 1/n$. We verify that the sequence $\{x_n\}$ is a Cauchy sequence: for any $\varepsilon > 0$ choose $m > 0$ so that $2\varepsilon_m < \varepsilon$. Then all the points x_k , for $k \geq m$, are in $B(x_m, \varepsilon_m)$. So for any $k, \ell \geq m$

$$d(x_k, x_\ell) \leq d(x_k, x_m) + d(x_m, x_\ell) < 2\varepsilon_m < \varepsilon.$$

Therefore the sequence converges to a point x in X . Clearly, $x \in \bigcap_{n=1}^{\infty} \overline{B}(x_n, \varepsilon_n)$. Finally we observe that by construction

$$\emptyset \neq \bigcap_{n=1}^{\infty} \overline{B}(x_n, \varepsilon_n) \subset \bigcap_{n=1}^{\infty} G_n \cap B.$$

□

As an application we prove a characterization of transitive maps.

Theorem 3.3.4. *Let (X, d) be a complete, separable, metric space without isolated points. Let $T : X \rightarrow X$ be a continuous transformation. Then the following are equivalent.*

- (1) *T is topologically transitive.*
- (2) *The set of points in X that have a dense positive orbit is a dense \mathcal{G}_δ set.*
- (3) *For all nonempty open sets U and V there exists an integer $n > 0$ such that $T^{-n}(U) \cap V \neq \emptyset$.*
- (4) *For all nonempty open sets U and V there exists an integer $n > 0$ such that $T^n(U) \cap V \neq \emptyset$.*

Proof. (1) \Rightarrow (4): Let x be a point with a dense positive orbit. Then there exists $k > 0$ so that $T^k(x) \in U$. Let $u = T^k(x)$. As X has no isolated points u has a dense positive orbit (Exercise 9). Then there exists $n > 0$ such that $T^n(u) \in V$. This means $T^n(U) \cap V \neq \emptyset$.

(4) \Rightarrow (3): There exists $n > 0$ so that $T^n(U) \cap V \neq \emptyset$. This means that there is an element of $T^n(U)$, which must have the form $T^n(u)$ for some $u \in U$, that is in V . So u is in $T^{-n}(V)$ and also in U . Thus, $U \cap T^{-n}(V) \neq \emptyset$.

(3) \Rightarrow (2): As X is separable, we can choose a sequence $\{x_m\}$ that is dense in X . Let $\{r_k\}$ be a countable sequence decreasing to 0

(say, $r_k = 1/k$). Note that if

$$x \in \bigcap_{m,k \geq 1} \bigcup_{n \geq 1} T^{-n}(B(x_m, r_k)),$$

then for all $m, k \geq 1$, $T^n(x) \in B(x_m, r_k)$ for some $n \geq 1$. This means that the positive orbit of x is dense. By the hypothesis, for all $m, k \geq 1$, any nonempty open set U must intersect $\bigcup_{n \geq 1} T^{-n}(B(x_m, r_k))$, so it follows that the set $\bigcup_{n \geq 1} T^{-n}(B(x_m, r_k))$ is dense. As the set $\bigcap_{m,k \geq 1} \bigcup_{n \geq 1} T^{-n}(B(x_m, r_k))$ is a countable intersection of dense open sets, the Baire category theorem implies that this set is dense. So the set of points with a dense orbit is dense.

Finally, it is clear that (2) \Rightarrow (1). □

Regarding completeness for metric spaces we observe that there can be two metrics that generate the same open sets for a space X but one is complete and the other is not.

Exercises

- (1) Let $T : [0, 1] \rightarrow [0, 1]$ be defined by $T(x) = 2x$ if $0 \leq x \leq 1/2$ and $T(x) = 2 - 2x$ if $1/2 < x \leq 1$. Show that T is finite measure-preserving. Find a point $x \in [0, 1)$ such that the (positive) orbit of x under T is dense.
- (2) For an integer $k > 1$ define $T_k(x) = kx \pmod{1}$ for $x \in [0, 1)$. Show that T_k is measure-preserving for Lebesgue measure. Find other measures in $[0, 1)$ that are invariant under T_k .
- (3) Let T be the doubling map. Show that the set of points that are periodic for T is a dense set in $[0, 1)$.
- (4) Let T be the doubling map. Find a point $x \in [0, 1)$ whose T -orbit is dense.
- (5) Show that the doubling map is continuous with respect to the metric d of Section 3.2.
- (6) Show that the doubling map is topologically transitive.
- (7) Let $f(x) = x^2$ be defined in $[0, 1]$. Is f topologically transitive? Show that it has infinitely many periodic points.
- (8) Show that the tent map T is topologically transitive.

-
- (9) Let $T : X \rightarrow X$ be a continuous map of a metric space with no isolated points and let $p \in X$. Show that if x is in the closure of the positive orbit of p , then x is an accumulation point of the positive orbit of p . Conclude that if p has a positive dense orbit, then every point in the positive orbit of p has a positive dense orbit.
- (10) Let X be a complete separable metric space with no isolated points and let $T : X \rightarrow X$ be a homeomorphism. Show that if there is a point $x \in X$ such that its orbit $\{T^n(x)\}_{n=-\infty}^{\infty}$ is dense, then T is (one-sided) topologically transitive.
- (11) Let (X, d) be a complete metric space. Show that T is topologically transitive if and only if for any closed set F such that $F \subset T^{-1}(F)$, then $F = X$ or F is nowhere dense.
- (12) Let A be a \mathcal{G}_δ subset of a metric space. Show that if A is dense, then its complement A^c is meager.
- (13) A set A in \mathbb{R} is said to be **residual** if its complement is a meager set. Show that the set of Liouville numbers is residual. Deduce that \mathbb{R} can be written as the disjoint union of a null set and a meager set.
- (14) Let X be a nonempty set. A **σ -ideal** on X is a collection of subsets of X that contains \emptyset and is closed under subsets and countable unions. Let (X, d) be a metric space. Show that the collection of meager sets in X is a σ -ideal.
- (15) Construct an open dense subset of $[0, 1]$ of arbitrarily small measure. (Hint: Let $\varepsilon > 0$ and let q_n be a countable dense subset in $[0, 1]$. Choose an open ball of radius $\varepsilon/2^n$ around each $q_n, n \geq 1$.)
- (16) Find all periodic points for the tent map.
- (17) Show that there cannot exist an invertible continuous map of the interval $[0, 1]$ that is topologically transitive.

3.4. Measure-Preserving Transformations

This section studies measure-preserving transformations in more detail. Let (X, \mathcal{S}, μ) be a measure space. We shall call a measure-preserving transformation $T : X \rightarrow X$ **finite measure-preserving** if $\mu(X) < \infty$ and **infinite measure-preserving** otherwise. When $\mu(X) = 1$, we may say that T is a **probability-preserving** transformation. In the infinite measure-preserving case we shall only consider the case when X is σ -finite. We say that (X, \mathcal{S}, μ, T) is a **measure-preserving dynamical system** if (X, \mathcal{S}, μ) is a σ -finite measure space and $T : X \rightarrow X$ is a measure-preserving transformation.

The main result is that to show that a transformation is measure-preserving it suffices to check the measure-preserving property on a sufficient semi-ring.

In applications, sufficient semi-rings will often be some collection of intervals such as the (left-closed, right-open) dyadic intervals. While the statement of the theorem is important, the proof may be omitted on a first reading.

Theorem 3.4.1. *Let (X, \mathcal{S}, μ) be a σ -finite measure space with a sufficient semi-ring \mathcal{C} . If for all I in \mathcal{C} ,*

- (1) $T^{-1}(I)$ is a measurable set, and
- (2) $\mu(T^{-1}(I)) = \mu(I)$,

then T is a measure-preserving transformation.

Proof. It suffices to show that for any $A \in \mathcal{S}(X)$ with $\mu(A) < \infty$, $T^{-1}(A)$ is measurable and $\mu(T^{-1}(A)) = \mu(A)$.

The proof consists of two parts. In the first part we show that for each measurable set A , if $H(A)$ is as defined in Lemma 2.7.2, then $T^{-1}(H(A))$ is measurable and $\mu(T^{-1}(H(A))) = \mu(H(A))$. We know that

$$H(A) = \bigcap_{n=1}^{\infty} H_n,$$

with $H_n \supset H_{n+1}$ and $\mu(H_n) < \infty$. Furthermore, each H_n can be written as

$$H_n = \bigsqcup_{i=1}^{\infty} C_{n,i},$$

with $C_{n,i} \in \mathcal{C}$. Thus, the set

$$T^{-1}(H_n) = T^{-1}\left(\bigsqcup_{i \geq 1} C_{n,i}\right) = \bigsqcup_{i \geq 1} T^{-1}(C_{n,i})$$

is measurable. Also,

$$\begin{aligned} \mu(T^{-1}(H_n)) &= \mu\left(\bigsqcup_{i \geq 1} T^{-1}(C_{n,i})\right) = \sum_{i \geq 1} \mu(T^{-1}(C_{n,i})) \\ &= \sum_{i \geq 1} \mu(C_{n,i}) = \mu\left(\bigsqcup_{n \geq 1} C_{n,i}\right) = \mu(H_n). \end{aligned}$$

Therefore $T^{-1}(H(A)) = T^{-1}\left(\bigcap_{n \geq 1} H_n\right) = \bigcap_{n \geq 1} T^{-1}(H_n)$ is measurable, and as $\mu(H_n) < \infty$, using Proposition 2.5.2,

$$\begin{aligned} \mu(T^{-1}(H(A))) &= \mu\left(\bigcap_{n \geq 1} T^{-1}(H_n)\right) = \lim_{n \rightarrow \infty} \mu(T^{-1}(H_n)) \\ &= \lim_{n \rightarrow \infty} \mu(H_n) = \mu\left(\bigcap_{n \geq 1} H_n\right) = \mu(H(A)). \end{aligned}$$

This concludes the proof of the first part.

For the second part of the proof we use that $A \subset H(A)$ and $\mu(H(A) \setminus A) = 0$. Write $N = H(A) \setminus A$.

Apply Lemma 2.7.2 again, this time to the null set N to obtain a null set $H(N)$ containing N . By the first part, $T^{-1}(H(N))$ is measurable and $\mu(T^{-1}(H(N))) = \mu(H(N)) = \mu(N) = 0$. As $N \subset H(N)$, then $T^{-1}(N) \subset T^{-1}(H(N))$, so $\mu(T^{-1}(N)) = 0$. Therefore $T^{-1}(N)$ is measurable.

Finally, note that

$$T^{-1}(A) = T^{-1}(H(A)) \setminus T^{-1}(N).$$

This shows that $T^{-1}(A)$ is measurable and, furthermore,

$$\begin{aligned} \mu(T^{-1}(A)) &= \mu(T^{-1}(H(A))) - \mu(T^{-1}(N)) \\ &= \mu(H(A)) - 0 = \mu(A). \end{aligned}$$

□

Second Proof: We give another shorter but nonconstructive proof of this theorem. Let

$$\mathcal{A} = \{A : A \in \mathcal{S} \text{ and } T^{-1}(A) \in \mathcal{S}, \mu(T^{-1}(A)) = \mu(A)\}.$$

Clearly \mathcal{A} contains the semi-ring \mathcal{C} . One can verify that \mathcal{A} is a monotone class. It follows that \mathcal{A} contains \mathcal{S} , so T is measure-preserving, completing the proof.

Example. We give another proof that the Doubling Map is measure-preserving. Let \mathcal{D} be the collection of left-closed right-open dyadic intervals in $[0, 1)$. For I in \mathcal{D} , write $I = [k/2^i, (k+1)/2^i)$ for integers k, i with $i \geq 0, k \in \{0, \dots, 2^i - 1\}$. Then

$$T^{-1}(I) = \left[\frac{k}{2^{i+1}}, \frac{k+1}{2^{i+1}} \right) \cup \left[\frac{k+2^i}{2^{i+1}}, \frac{k+1+2^i}{2^{i+1}} \right).$$

Evidently, $T^{-1}(I)$ is measurable and one can check that $\mu(T^{-1}(I)) = \frac{1}{2^i} = \mu(I)$. We saw in Section 2.7 that \mathcal{D} is a sufficient semi-ring, so an application of Theorem 3.4.1 yields that T is measure-preserving.

Exercises

- (1) (Boole's transformation) Let $T : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $T(x) = x - \frac{1}{x}$ if $x \neq 0$ and $T(0) = 0$. Show that T is measure-preserving on \mathbb{R} with Lebesgue measure.
- (2) Show that if (X, \mathcal{S}, μ) is a σ -finite measure-space and $T : X \rightarrow X$ is measure-preserving, then for any $X_0 \in \mathcal{S}(X)$ with $T^{-1}(X_0) = X_0$, the system $(X_0, \mathcal{S}(X_0), \mu, T)$ is a measure-preserving dynamical system.
- (3) Let (X, \mathcal{S}, μ) be a σ -finite measure space and let $X_0 \in \mathcal{S}(X)$ with $\mu(X \setminus X_0) = 0$. Suppose there exists a transformation T_0 so that $(X_0, \mathcal{S}(X_0), \mu, T_0)$ is a measure-preserving dynamical system. Show that there exists a transformation $T : X \rightarrow X$ so that $T(x) = T_0(x)$ for $x \in X_0$ and (X, \mathcal{S}, μ, T) is a measure-preserving dynamical system. (T is not unique but differs from T_0 on only a null set.)
- (4) Show that if (X, \mathcal{S}, μ, T) is a measure-preserving dynamical system, then for any integer $n > 0$, $(X, \mathcal{S}, \mu, T^n)$ is a measure-preserving dynamical system.

- (5) Show that if (X, \mathcal{S}, μ, T) is an invertible measure-preserving dynamical system, then for any integer n , $(X, \mathcal{S}, \mu, T^n)$ is an invertible measure-preserving dynamical system.
- (6) Complete the details of the second proof of Theorem 3.4.1.

3.5. Recurrence

A measure-preserving transformation T defined on a measure space (X, \mathcal{S}, μ) is said to be **recurrent** if for every measurable set A of positive measure there is a null set $N \subset A$ such that for all $x \in A \setminus N$ there is an integer $n = n(x) > 0$ with

$$T^n(x) \in A.$$

Informally, T is recurrent when for any set A of positive measure almost every point of A returns to A at some future time. We think of n as a “return time” to A . Figure 3.7 shows a typical point in a set of positive measure for a recurrent transformation. We will see in Theorem 3.5.3 that every finite measure-preserving transformation is recurrent. However, the transformation $T : [0, \infty) \rightarrow [0, \infty)$ defined by $T(x) = x + 1$ is measure-preserving for Lebesgue measure, but is easily seen not to be recurrent (let $A = [0, 1)$, for example).

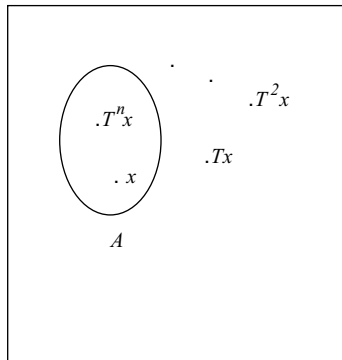


Figure 3.7. T is recurrent

As the following lemma shows, once a transformation is recurrent there exist infinitely many return times.

Lemma 3.5.1. *Let (X, \mathcal{S}, μ) be a measure space and $T : X \rightarrow X$ a recurrent measure-preserving transformation. Then, for all sets of positive measure A , there exists a null set N such that for all $x \in A \setminus N$ there is an increasing sequence $m_i > 0$ with $T^{m_i}(x) \in A \setminus N$ for all $i \geq 1$.*

Proof. By definition, there is a null set N_1 such that for all $x \in A \setminus N_1$ there is an integer $n = n(x)$ with $T^n(x) \in A$. Let $N = \bigcup_{k=0}^{\infty} T^{-k}(N_1)$. N is a null set as $\mu(T^{-k}(N_1)) = \mu(N_1) = 0$ for $k \geq 0$. For all $k \geq 0$, $T^{-k}(N) \subset N$, so if $x \notin N$, then $T^k(x) \notin N$. Thus, if we choose $x \in A \setminus N \subset A \setminus N_1$, then $T^{n(x)}(x) \in A \setminus N$. Let $m_1 = n(x)$. Applying the result we just proved to the point $z = T^{m_1}(x)$ that is in $A \setminus N$, we obtain an integer $n_1 = n(z) > 0$ so that $T^{n_1}(z) \in A \setminus N$. If we let $m_2 = m_1 + n_1$, then $T^{m_2}(x) = T^{n_1}(z) \in A \setminus N$. Continue in this manner to obtain an increasing sequence of integers m_i . \square

The following lemma shows a useful characterization of recurrence. A measure-preserving transformation T on X is said to be **conservative** if for any set A of positive measure there exists an integer $n > 0$ such that $\mu(A \cap T^{-n}(A)) > 0$.

Lemma 3.5.2. *Let (X, \mathcal{S}, μ, T) be a measure space. A measure-preserving transformation T on X is recurrent if and only if it is conservative.*

Proof. First observe that T is recurrent if and only if

$$\mu\left(A \setminus \bigcup_{n=1}^{\infty} T^{-n}(A)\right) = 0$$

for all sets A of positive measure. This is so since the set $\bigcup_{k=1}^{\infty} T^{-k}(A)$ consists of all points that are in A after some positive iteration, and recurrence is precisely the statement that the set of points in A that are not in A after some positive iteration is a set of measure zero.

Now suppose that T is recurrent and let A be a set of positive measure. Then, $\mu\left(A \setminus \bigcup_{k=1}^{\infty} [A \cap T^{-k}(A)]\right) = \mu\left(A \setminus \bigcup_{k=1}^{\infty} T^{-k}(A)\right) = 0$. As A has positive measure, it must be the case that for some integer $n > 0$, $\mu(A \cap T^{-n}(A)) > 0$.

For the converse let A be a set of positive measure. Set

$$B = A \setminus \bigcup_{k=1}^{\infty} T^{-k}(A).$$

If $\mu(B) > 0$, then there is an integer $n > 0$ such that $\mu(B \cap T^{-n}(B)) > 0$. This would mean that there is a point $x \in B$ such that $T^n(x) \in B$, contradicting the definition of B . Therefore $\mu(B) = 0$ and T is recurrent. \square

The following is probably the first theorem in ergodic theory; it was shown by Poincaré in 1899 and used in his study of celestial mechanics.

Theorem 3.5.3 (Poincaré Recurrence). *Let (X, \mathcal{S}, μ) be a finite measure space. If $T : X \rightarrow X$ is a measure-preserving transformation, then T is a recurrent transformation.*

Proof. By Lemma 3.5.2 it suffices to show that for any set of positive measure A , there is an integer $n > 0$ such that $\mu(A \cap T^{-n}(A)) > 0$. Suppose, on the contrary, that $\mu(A \cap T^{-n}(A)) = 0$ for all $n > 0$. Then for any nonnegative integers $k \neq \ell$, writing $\ell = n + k, n > 0$,

$$\begin{aligned} \mu(T^{-\ell}(A) \cap T^{-k}A) &= \mu(T^{-n-k}(A) \cap T^{-k}(A)) \\ &= \mu(T^{-k}[T^{-n}(A) \cap A]) = \mu(T^{-n}(A) \cap A) = 0. \end{aligned}$$

This means that the sets $\{T^{-n}(A)\}_{n \geq 0}$ are almost pairwise disjoint, so (see Exercise 2.5.2)

$$\mu\left(\bigcup_{n=0}^{\infty} T^{-n}(A)\right) = \sum_{n=0}^{\infty} \mu(T^{-n}(A)) = \sum_{n=0}^{\infty} \mu(A) = \infty,$$

a contradiction as $\mu(X) < \infty$. Thus it must be that $\mu(A \cap T^{-n}(A)) > 0$ for some $n > 0$. \square

We present another useful characterization of recurrence. A set C is said to be **compressible** for a measure-preserving transformation T if $T^{-1}(C) \subset C$ and $\mu(C \setminus T^{-1}(C)) > 0$. T is **incompressible** if it admits no compressible sets.

Lemma 3.5.4. *Let (X, \mathcal{S}, μ, T) be a σ -finite measure space. A measure-preserving transformation T on X is recurrent if and only if it is incompressible.*

Proof. Suppose T is recurrent and let C be a set such $T^{-1}(C) \subset C$. Put $A = C \setminus T^{-1}(C)$. We claim that

$$T^{-n}(A) \cap A = \emptyset \text{ for all } n \geq 1.$$

For suppose that $x \in A$. Then $x \in C$ and $T(x) \notin C$. So $T(x) \notin T^{-1}(C)$, or $T^2(x) \notin C$; in this way one obtains that $T^n(x) \notin C$ for all $n \geq 1$. Thus $T^n(x) \notin A$ and the claim is proved. Lemma 3.5.2 implies that $\mu(A) = 0$, so T is incompressible.

Now let T be incompressible and let A be a set of positive measure. Put $C = \bigcup_{i=0}^{\infty} T^{-i}(A)$. Then $T^{-1}(C) \subset C$. One can verify that

$$C \setminus T^{-1}(C) = A \setminus \bigcup_{i=1}^{\infty} T^{-i}(A).$$

As T is incompressible, $\mu(A \setminus \bigcup_{i=1}^{\infty} T^{-i}(A)) = 0$, which implies that T is recurrent. \square

Recurrence was generalized in a significant way in 1977 by Furstenberg. A measure-preserving transformation of a probability space (X, \mathcal{S}, μ) is said to be **multiply recurrent** if for all sets A of positive measure and for each integer $k > 0$ there exists an integer $n > 0$ so that

$$\mu(A \cap T^{-n}(A) \cap T^{-2n}(A) \cap \dots \cap T^{-(k-1)n}(A)) > 0.$$

The following theorem is due to Furstenberg. Its proof, which we omit, is significantly harder than the proof of Poincaré recurrence.

Theorem 3.5.5 (Furstenberg Multiple Recurrence). *If T is a measure-preserving transformation on a finite Lebesgue measure space (X, \mathcal{S}, μ) , then T is a multiply recurrent transformation.*

For a proof of this theorem, which uses techniques beyond those developed in this book, the reader is referred to [24]. Furstenberg used the Multiple Recurrence theorem to give a dramatically new proof of a deep result in number theory called Szemerédi's theorem. Szemerédi's

theorem asserts that a set of positive density in the integers (for example, the even numbers have density $1/2$) contains arithmetic progressions of arbitrary length.

Exercises

- (1) Let $T : X \rightarrow X$ be a finite measure-preserving transformation. Given a measurable set A of positive measure, let n be the first integer such that $\mu(T^{-n}(A) \cap A) > 0$. Find the best upper bound (in terms of the measure of A and X) for n . Prove your claim.
- (2) Let A be the set of all $x \in [0, 1]$ with the following property: if $0.x_1x_2 \dots x_k \dots$ is the decimal expansion of x , then for each integer $k > 0$, the string $.x_1x_2 \dots x_k$ appears infinitely often in the sequence $x_1x_2 \dots x_k \dots$ of the decimal expansion of x . Show that $[0, 1] \setminus A$ is a null set.
 Show that there is a null set N so that for all $x \in [0, 1] \setminus N$, for all integers $k > 0$, if the decimal expansion of x starts with $x = 0.x_1x_2 \dots x_k \dots$, then the string $x_1x_2 \dots x_k$ appears infinitely often in the decimal expansion of x .
- (3) Show that an irrational rotation is multiply recurrent without using Theorem 3.5.5.
- (4) A set $P \subset \mathbb{N}$ is said to be a **Poincaré sequence** if for every finite measure-preserving system (X, \mathcal{S}, μ, T) and any set $A \in \mathcal{S}$ of positive measure there exists $n \in P, n \neq 0$, such that $\mu(T^{-n}(A) \cap A) > 0$. Show that for any infinite set P the set of differences $P - P$ (consisting of points of the form $a - b$ where $a, b \in P$) is a Poincaré sequence.
- (5) A set $Q \subset \mathbb{N}$ is said to be a **thick set** if it contains intervals of integers of arbitrary length. Show that a thick set is a Poincaré sequence.
- (6) A measurable set W is said to be **wandering** for a measure-preserving transformation T if for all $i, j \geq 0$ with $i \neq j$, $T^{-i}(W) \cap T^{-j}(W) = \emptyset$. Show that T is recurrent if and only if it has no wandering sets of positive measure.

3.6. Almost Everywhere and Invariant Sets

The principal notions in measure theory and ergodic theory remain invariant under a change by a set of measure zero. For example, it follows from Lemma 4.2.6 that if $T : X \rightarrow X$ is a measurable transformation and $S : X \rightarrow X$ is a transformation that differs from T on a null set, i.e., $\mu(\{x : T(x) \neq S(x)\}) = 0$, then S is also a measurable transformation. We shall also see in Lemma 4.6.2 that the Lebesgue integral of a function does not change if the function is changed on a set of measure zero. In addition, there are several important theorems, such as the Ergodic Theorem, that hold only after discarding a set of measure zero. The idea of carefully discarding a set of measure zero is a fundamental idea in measure theory and ergodic theory. Typically, one only needs to define objects up to a set of measure zero. We shall say that an equality holds **almost everywhere**, written as **a.e.**, if it holds outside a null set.

We start by generalizing some of the concepts we have already defined to allow changes on a null set. The first one is the notion of an invertible measurable transformation. According to the principle of ignoring things that happen on a null set, if a transformation fails to be invertible on a null set, it should still be considered invertible. So one is led to define the notion of an invertible transformation mod μ as a transformation that is invertible after discarding a μ -measure zero set from its domain. However, the remaining map should be a transformation, i.e., points that are not in the null set should also miss the null set under iteration by the transformation. More precisely, if $T : X \rightarrow X$ is a transformation and $N \subset X$ is the null set that is discarded, it is necessary that if $x \in X \setminus N$, then $T(x) \in X \setminus N$. To express this idea in a clear way we are led to the notion of an *invariant set*, which plays a crucial role in dynamics.

Let $T : X \rightarrow X$ be a transformation. A subset $A \subset X$ is said to be **positively invariant** or **positively T -invariant** when

$$x \in A \text{ implies } T(x) \in A.$$

Equivalently, $A \subset T^{-1}(A)$. Then, the transformation T restricted to A defines a transformation $T : A \rightarrow A$. It is also clear that if T restricted to A defines a transformation, then A is a positively

invariant set. In some cases we need an additional property: A set A is **strictly T -invariant** when $T^{-1}(A) = A$, i.e., when $x \in A$ if and only if $T(x) \in A$. In ergodic theory we often call a strictly invariant set simply **invariant** or **T -invariant**.

Example. Let $X = [0, \infty)$ and define $T : X \rightarrow X$ by $T(x) = x + 1$, an invertible measure-preserving transformation. Then $A = [1, \infty)$ is positively invariant but not strictly invariant.

Question. Let $T : X \rightarrow X$ be a transformation. Show that the empty set and the whole space X are positively invariant sets. Show that X is strictly invariant if and only if T is onto.

Example. Let $X = \{0, 1, 2, \dots, 5\}$. Define $T : X \rightarrow X$ by $T(0) = 1, T(1) = 2, T(2) = 0, T(3) = 4, T(4) = 5, T(5) = 3$, and $S(i) = i + 1$ for $i = 0, \dots, 4, S(5) = 0$. Let $A = \{0, 1, 2\}$. Then A is a strictly T -invariant set but is not a positively S -invariant set. T restricted to A is a rotation on three points. One can also verify that S does not have any positively invariant sets other than \emptyset and X .

Given two sets $A, B \subset X$, we say that

$$A = B \text{ mod } \mu \text{ if } \mu(A \triangle B) = 0.$$

(Recall that $A \triangle B = \emptyset$ if and only if $A = B$.) A set A is said to be **strictly invariant mod μ** (or **T -invariant mod μ**) if $A = T^{-1}(A) \text{ mod } \mu$.

Question. Let (X, \mathcal{S}, μ) be a measure space and $T : X \rightarrow X$ a transformation. Show that if T is measure-preserving, then X is strictly invariant mod μ .

The following lemma is an immediate consequence of Lemma 3.5.4 but it is important to keep in mind.

Lemma 3.6.1. *Let T be a recurrent measure-preserving transformation on a measure space (X, \mathcal{S}, μ) . If A is positively invariant, then it is strictly invariant mod μ .*

Furthermore, Lemma 3.6.2 shows that a strictly invariant mod μ set differs from a strictly invariant set in a null set. Therefore, in the case of finite measure-preserving transformations, as we identify sets

that differ in a null set, a positively invariant set may be replaced by a strictly invariant set; this means that there is no essential distinction between positively invariant sets and strictly invariant sets. We shall see some examples from topological dynamics and infinite measure-preserving transformations where the distinction is important. For example, it can be shown that the transformation $T : \mathbb{Z} \rightarrow \mathbb{Z}$ defined on the integers with counting measure by $T(n) = n + 1$ is an infinite measure-preserving transformation that has no strictly invariant sets of positive measure other than \mathbb{Z} , but the set $A = \{n \in \mathbb{Z} : n > 0\}$ is an invariant set for T . There are also examples on nonatomic spaces when T is not invertible.

Lemma 3.6.2. *Let (X, \mathcal{S}, μ) be a σ -finite measure space and let $T : X \rightarrow X$ be a measure-preserving transformation. If $A \in \mathcal{S}$ is strictly T -invariant mod μ , then there exists a set \hat{A} that differs from A by a null set (i.e., $\mu(A \Delta \hat{A}) = 0$) that is strictly T -invariant.*

Proof. We assume that μ is finite and leave the σ -finite case to the reader. As A is strictly T -invariant mod μ , $\mu(T^{-1}(A) \Delta A) = 0$. By the triangle inequality (see Exercise 2.5.6) and the fact that T is measure-preserving,

$$\mu(T^{-2}(A) \Delta A) \leq \mu(T^{-2}(A) \Delta T^{-1}(A)) + \mu(T^{-1}(A) \Delta A) = 0.$$

So by induction,

$$\mu(T^{-n}(A) \Delta A) = 0 \text{ for all } n \geq 0.$$

By Exercise 6, the set $\hat{A} = \bigcap_{k=1}^{\infty} T^{-k}(A^+)$, where $A^+ = \bigcup_{n=1}^{\infty} T^{-n}(A)$, is strictly T -invariant. Now

$$\begin{aligned} \mu(A \Delta T^{-k}(A^+)) &= \mu(A \Delta \bigcup_{n=k+1}^{\infty} T^{-n}(A)) \\ &\leq \sum_{n=k+1}^{\infty} \mu(A \Delta T^{-n}(A)) = 0. \end{aligned}$$

Then by Theorem 2.5.2b) (here is where we use that μ is finite),

$$\mu(A \Delta \bigcap_{k=1}^{\infty} T^{-k}(A^+)) = 0,$$

so $\mu(A \Delta \hat{A}) = 0$. □

We conclude by applying the principle of neglecting sets of measure zero to extend the notion of an invertible transformation. A transformation $T : X \rightarrow X$ is said to be an **invertible measurable transformation** mod μ (or mod 0 when the measure is clear from the context) if there exists a measurable set $X_0 \subset X$, with $\mu(X \setminus X_0) = 0$ and such that T is one-to-one on X_0 (i.e., for any x_1, x_2 in X_0 , if $T(x_1) = T(x_2)$, then $x_1 = x_2$), $T^{-1}(X_0) = X_0$ and $T^{-1} : X_0 \rightarrow X_0$ is a measurable transformation (it follows that T^{-1} is measure-preserving). In the literature one may see authors refer to “invertible transformations” when what is really meant is “invertible transformations mod μ .” The idea here is that after discarding a set of measure zero one obtains an invertible transformation.

Exercises

- (1) Let T be an invertible mod μ measurable transformation. Show that T is measure-preserving if and only if for all measurable sets A , $T(A)$ is measurable and $\mu(T(A)) = \mu(A)$.
- (2) Show that if (X, \mathcal{S}, μ, T) is a measure-preserving dynamical system, then for any integer $n > 0$, $(X, \mathcal{S}, \mu, T^n)$ is a measure-preserving dynamical system.
- (3) Show that if (X, \mathcal{S}, μ, T) is an invertible measure-preserving dynamical system, then for any integer n , $(X, \mathcal{S}, \mu, T^n)$ is an invertible measure-preserving dynamical system.
- (4) Complete the proof of Lemma 3.6.2 in the σ -finite case.

We define some sets that are useful in discussing invariance properties and are used in the following exercises. Given a transformation $T : X \rightarrow X$ and a set $A \subset X$, define the sets

$$A^+ = \bigcup_{n=1}^{\infty} T^{-n}(A); \quad A^\oplus = \bigcup_{n=0}^{\infty} T^{-n}(A).$$

A^+ represents the set of points in X that at some time $n > 0$ enter A . For example, $\mu(A^+ \cap A) > 0$ means that there is a set of positive measure of points of A that “come back” to A at some positive time. The set A^\oplus is precisely A^+ with the addition of A . We note, however, that A^+ is not positively

invariant, but as $T^{-1}(A^+) \subset A^+$, it can be shown that the set

$$\widehat{A} = \bigcap_{n=1}^{\infty} T^{-n}(A^+)$$

is strictly invariant. When T is invertible it is often more convenient to use the set

$$A^* = \bigcup_{n=-\infty}^{\infty} T^{-n}(A).$$

- (5) Let $T : X \rightarrow X$ be a measure-preserving transformation and let $A \subset X$. Show that if N is a null set, then N^\oplus is a null set and $X \setminus N^\oplus$ is a positively T -invariant set, of the same measure as X , which does not contain N . Furthermore, if $T : X \rightarrow X$ is measure-preserving, then $T : X \setminus N^\oplus \rightarrow X \setminus N^\oplus$ is measure-preserving.
- (6) Show that the set \widehat{A} is strictly T -invariant and satisfies

$$\widehat{A} = \bigcap_{n=k}^{\infty} T^{-n}(A^+) = \bigcap_{n=k}^{\infty} T^{-n}(A^\oplus)$$

for all $k > 0$. If A has measure zero, then \widehat{A} has measure zero. Furthermore, if T is measure-preserving and A has finite measure, then $\mu(A) = \mu(\widehat{A})$.

- (7) Show that if T is invertible, then A^* is a strictly T -invariant set containing A . If furthermore N is a null set, then $N^* \supset N$ is a null set and $T : X \setminus N^* \rightarrow X \setminus N^*$ is an invertible transformation, and if $T : X \rightarrow X$ is measure-preserving, then $T : X \setminus N^* \rightarrow X \setminus N^*$ is measure-preserving.

3.7. Ergodic Transformations

Ergodicity is one of our most important concepts. Originally called *metric transitivity*, it was introduced by Birkhoff and Smith in 1928. We will see that there are several equivalent formulations of ergodicity, each having its own interesting interpretation. One of the equivalences is surprising and the result of a deep theorem, namely the

ergodic theorem (Theorem 5.1.1). We will discuss that interpretation later. In this section we study the basic definition of ergodicity and show that irrational rotations and the doubling map are ergodic transformations.

We are interested in studying the measurable dynamics of a transformation T defined on a space X . If there existed a set $A \subset X$ such that $x \in A$ if and only if $T(x) \in A$, then T restricted to A (i.e., as a map $T : A \rightarrow A$) and T restricted to A^c would both be dynamical systems in their own right, and it is reasonable to think that the study of the dynamics of T could be reduced to the study of the dynamics of T restricted to A and T restricted to A^c . Thus one can think of systems not having such strictly invariant sets A (other than \emptyset and X) as “indecomposable” or basic systems.

A measure-preserving transformation T is said to be **ergodic** if whenever A is a strictly T -invariant measurable set, then either $\mu(A) = 0$ or $\mu(A^c) = 0$. The following is the simplest example of an ergodic transformation.

Example. Let (X, \mathcal{S}, μ) be a one-point canonical Lebesgue space (i.e., $X = \{p\}$, $\mathcal{S} = \{\emptyset, X\}$ and μ is the counting measure on X defined by $\mu(\{p\}) = 1$). Let T be the identity transformation on X (i.e., $T(p) = p$). Then T is a measure-preserving ergodic transformation. To see that it is ergodic note that the strictly invariant sets are \emptyset and $\{p\}$, and they have measure 0 or their complement has measure 0.

Question. Let T be the identity transformation on a canonical Lebesgue measure space X . Show that T is ergodic if and only if X is a one-point space.

Question. Show that the transformation $R(x) = x + 3/4 \pmod{1}$ is a finite measure-preserving transformation on $[0, 1)$ that is not ergodic.

The following lemma shows that in the definition of ergodicity one may consider strictly invariant sets up to measure and not just strictly invariant.

Lemma 3.7.1. *Let (X, \mathcal{S}, μ) be a measure space and let $T : X \rightarrow X$ be a measure-preserving transformation. Then T is ergodic if and only if when A is strictly invariant mod μ , then $\mu(A) = 1$ or $\mu(A^c) = 0$.*

Proof. This is a direct consequence of Lemma 3.6.2. □

The transformation $T : \mathbb{Z} \rightarrow \mathbb{Z}$ defined by $T(n) = n + 1$, with counting measure on \mathbb{Z} , is ergodic but not recurrent. As we have seen, recurrence always holds in the finite measure-preserving case, and most interesting characterizations of ergodicity are in the recurrent case.

Lemma 3.7.2. *Let (X, \mathcal{S}, μ) be a σ -finite measure space and let $T : X \rightarrow X$ be a measure-preserving transformation. Then the following are equivalent.*

- (1) T is recurrent and ergodic.
- (2) For every set A of positive measure, $\mu(X \setminus \bigcup_{n=1}^{\infty} T^{-n}(A)) = 0$. (In this case we will say A **sweeps out** X .)
- (3) For every measurable set A of positive measure and for a.e. $x \in X$ there exists an integer $n > 0$ such that

$$T^n(x) \in A.$$

(This is illustrated in Figure 3.8.)

- (4) If A and B are sets of positive measure, then there exists an integer $n > 0$ such that

$$T^{-n}(A) \cap B \neq \emptyset.$$

- (5) If A and B are sets of positive measure, then there exists an integer $n > 0$ such that

$$\mu(T^{-n}(A) \cap B) > 0.$$

Proof. First we show that (5) implies (1). Let A be a strictly invariant set of positive measure. Then $T^{-n}(A) = A$ for all $n > 0$. Let $B = A^c$. If B had positive measure, then there would exist an integer $n > 0$ so that $\mu(T^{-n}(A) \cap B) > 0$, so $\mu(A \cap A^c) > 0$, a contradiction. Therefore, $\mu(A^c) = 0$, which means that T is ergodic.

Suppose (1) holds. Let A be a set of positive measure and set $A^+ = \bigcup_{n=1}^{\infty} T^{-n}(A)$. Then $T^{-1}(A^+) \subset A^+$; since T is recurrent $\mu(A^+ \setminus T^{-1}(A^+)) = 0$. Therefore A^+ is strictly invariant mod μ and since it has positive measure the ergodicity of T implies that $A^+ = X \text{ mod } \mu$.

It is clear that (2) is equivalent to (3). Now assume (2) and let A and B be sets of positive measure. Then $A^+ = X \text{ mod } \mu$ and so there is $n > 0$ with $\mu(T^{-n}(A) \cap B) > 0$. Therefore (4) holds.

Assume (4). Let A, B be sets of positive measure and suppose $\mu(T^{-n}(A) \cap B) = 0$ for all $n > 0$. Then put $A_0 = A \setminus \bigcup_{n=1}^{\infty} T^{-n}(A) \cap B$. Then $\mu(A_0) > 0$ but $T^{-n}(A_0) \cap B = \emptyset$, contradicting (4).

Now if (5) is true and A is strictly T -invariant (then $T^{-n}(A) = A$ for all $n > 0$), using $B = A^c$ in (5) we see that A and A^c cannot both have positive measure. Therefore, T is ergodic. Recurrence follows by letting $B = A$. \square

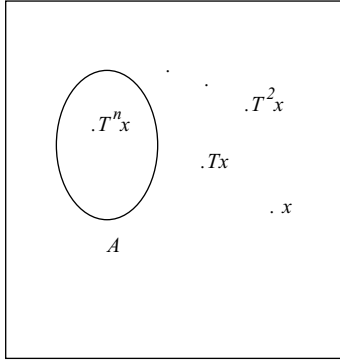


Figure 3.8. T is recurrent and ergodic

We now introduce some approximation techniques that will prove useful in showing ergodicity, and apply them to the case of irrational rotations and the doubling map.

Given measurable sets A and I (I will typically be in some sufficient semi-ring, i.e., an interval or a dyadic interval) and given $1 > \delta > 0$ (usually chosen small), we say that I is $(1 - \delta)$ -**full** of

A if

$$\lambda(A \cap I) > (1 - \delta)\lambda(I).$$

While we say “ $(1 - \delta)$ -full,” strictly speaking we should say “more than $(1 - \delta)$ -full.” If $\delta = 1/4$, say, then “ I is $3/4$ -full of A ” can be interpreted as saying that “in measure, more than $3/4$ of I is in A .”

The following lemma shows that for any set of positive finite measure and any $\delta > 0$ there is always an element of a sufficient semi-ring that is $(1 - \delta)$ -full of the set.

Lemma 3.7.3. *Let $(X, \mathfrak{L}, \lambda)$ be a nonatomic measure space with a sufficient semi-ring \mathcal{C} . If $A \in \mathfrak{L}$ is of finite positive measure, then for any $\delta > 0$ there exists $I \in \mathcal{C}$ such that I is $(1 - \delta)$ -full of A .*

Proof. Let $\varepsilon > 0$ be such that $\varepsilon < \frac{\delta}{2 - \delta}$ (this choice of ε will be apparent later in the proof). By Lemma 2.7.3 there exists $H^* = \bigsqcup_{j=1}^N I_j$, with the sets I_j disjoint and in \mathcal{C} , such that $\lambda(A \Delta H^*) < \varepsilon\lambda(A)$. By Exercise 2.5.4,

$$\lambda(A \cap H^*) > (1 - \varepsilon)\lambda(A).$$

Also, by Exercise 2.5.5,

$$\lambda(H^*) < \lambda(A) + \varepsilon\lambda(A) = (1 + \varepsilon)\lambda(A).$$

Now assume that for all $j \in \{1, \dots, N\}$,

$$\lambda(A \cap I_j) \leq (1 - \delta)\lambda(I_j).$$

Then $\lambda(A \cap H^*) \leq (1 - \delta)\lambda(H^*)$. Therefore,

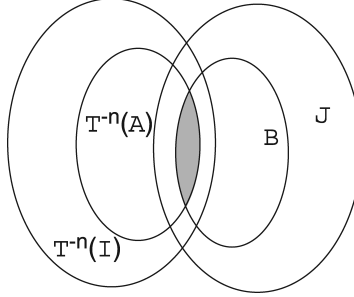
$$\lambda(A \cap H^*) < (1 - \delta)(1 + \varepsilon)\lambda(A).$$

Now ε was chosen so that $(1 - \delta)(1 + \varepsilon) < 1 - \varepsilon$, but this means that $\lambda(A \cap H^*) < \lambda(A \cap H^*)$, a contradiction. Therefore there must exist $j \in \{1, \dots, N\}$ such that $\lambda(A \cap I_j) > (1 - \delta)\lambda(I_j)$. \square

The following lemma contains a basic inequality that will be used in several proofs of ergodicity. It follows directly from set theory.

Lemma 3.7.4. *Let T be a measure-preserving transformation and let A, B, I, J be any measurable sets such that $A \subset I$ and $B \subset J$. Then for all integers $n \geq 0$,*

$$(3.6) \quad \lambda(T^{-n}(A) \cap B) \geq \lambda(T^{-n}(I) \cap J) - \lambda(I \setminus A) - \lambda(J \setminus B).$$

Figure 3.9. $T^{-n}(A) \cap B$

Proof. Note that

$$T^{-n}(I) \cap J \subset (T^{-n}(A) \cap B) \cup (T^{-n}(I) \setminus T^{-n}(A)) \cup (J \setminus B),$$

and $\lambda(T^{-n}(I) \setminus T^{-n}(A)) = \lambda(I \setminus A)$. (See Figure 3.9.) \square

We are now ready for our main result.

Theorem 3.7.5. *Irrational rotations are ergodic.*

Proof. Let R be the rotation by the irrational number α on $X = [0, 1)$. Let A_1 and B_1 be any sets of positive measure. Then, there exist dyadic intervals I and J such that

$$\lambda(A_1 \cap I) > \frac{3}{4}\lambda(I) \quad \text{and} \quad \lambda(B_1 \cap J) > \frac{3}{4}\lambda(J).$$

Furthermore, we may assume that I and J are of the same measure (if J , say, is bigger than I , then at least one of the two halves of J must be $\frac{3}{4}$ -full of B_1 ; continue in this way until obtaining a subinterval of J of the same measure as I that is $\frac{3}{4}$ -full of B_1 , finally rename it J). Write

$$A = A_1 \cap I \quad \text{and} \quad B = B_1 \cap J.$$

Suppose $I = [a, b)$ is to the left of $J = [c, d)$ in $[0, 1)$, i.e., $a \leq c$. As the orbit of b under R is dense, there is an integer $n > 0$ such that

$$d - \frac{d-c}{4} < R^n(b) < d.$$

Therefore $\lambda(R^n(I) \cap J) > \frac{3}{4}\lambda(J)$. Thus by Lemma 3.7.4,

$$\begin{aligned}\lambda(R^n(A) \cap B) &\geq \lambda(R^n(I) \cap J) - \lambda(I \setminus A) - \lambda(J \setminus B) \\ &> \frac{3}{4}\lambda(J) - \frac{1}{4}\lambda(I) - \frac{1}{4}\lambda(J) > 0.\end{aligned}$$

Therefore R is ergodic. \square

A measure-preserving transformation T is said to be **totally ergodic** if for all integers $n > 0$, T^n is ergodic. Irrational rotations are our first example of totally ergodic transformations as $R_\alpha^n = R_{n\alpha}$ and so R_α^n is ergodic for all $n \neq 0$ if α is irrational.

We end with an example of a noninvertible ergodic transformation.

Theorem 3.7.6. *Let $T(x) = 2x \pmod{1}$ be defined on $[0, 1)$. Then T is an ergodic finite measure-preserving transformation on $[0, 1)$.*

Proof. We have already seen that T is finite measure-preserving.

To show ergodicity, first let $D_{n,k} = [\frac{k}{2^n}, \frac{k+1}{2^n})$ ($n > 0$, $k = 0, \dots, 2^n - 1$) be a dyadic interval in $[0, 1)$. We note that $T^n(D_{n,k}) = [0, 1)$. Using this, it follows by induction that $T^{-n}(D_{n,k})$ consists of 2^n disjoint dyadic intervals each of length 2^{-2n} . In Exercise 4, the reader is asked to show by induction on n that for any measurable set A ,

$$\lambda(T^{-n}(A) \cap D_{n,k}) = \frac{1}{2^n}\lambda(A) = \lambda(A)\lambda(D_{n,k})$$

for $k = 0, \dots, 2^n - 1$. Suppose now that A is a strictly T -invariant set. So $T^{-n}(A) = A$ for all $n > 0$, and

$$\lambda(A \cap D_{n,k}) = \lambda(A)\lambda(D_{n,k}).$$

If A has positive measure, as the dyadic intervals form a sufficient semi-ring, for any $\delta > 0$ there exists a dyadic interval $D_{n,k}$ (for some n, k) so that $\lambda(A \cap D_{n,k}) > (1 - \delta)\lambda(D_{n,k})$. As δ is arbitrary, this implies that $\lambda(A) = 1$ and therefore T is ergodic. \square

Exercises

- (1) Show that if R is an irrational rotation, then R^n is ergodic for all $n \neq 0$.

- (2) Let T be a rotation on n points, i.e., $X = \{a_0, a_1, \dots, a_{n-1}\}$, μ a measure on subsets of X defined by $\mu(\{a_i\}) = 1/n$, and T a transformation on X defined by $T(a_i) = a_{i+1 \pmod n}$. Show that T is ergodic.
- (3) Let T be a totally ergodic measure-preserving transformation. Show that if T is invertible, then T^n is ergodic for all $n < 0$.
- (4) Complete the details in the proof of Theorem 3.7.6.
- (5) Show that for each integer $k > 1$ the transformation $T(x) = kx \pmod 1$ is ergodic on $[0, 1)$.
- (6) Show that the 2-dimensional baker's transformation is an invertible (mod λ) measure-preserving ergodic transformation.
- (7) Let T be the doubling map of Theorem 3.7.6. Is T totally ergodic?

3.8. The Dyadic Odometer

The transformation studied in this section was defined by Kakutani and von Neumann in the 1940's and is also called the Kakutani–von Neumann odometer or the dyadic adding machine. We first present its definition as a piecewise-linear map on infinitely many pieces in the unit interval. While this defines the transformation on the unit interval and helps understand why it is measure-preserving, there is another definition that is better for discussing its dynamical properties, such as ergodicity. This other definition also introduces a new method for constructing transformations, called *cutting and stacking*, which will provide us with a wealth of examples and counterexamples.

We start by defining $T : [0, 1) \rightarrow [0, 1)$ by

$$T(x) = \begin{cases} x + \frac{1}{2}, & \text{if } 0 \leq x < 1/2; \\ x - \frac{1}{4}, & \text{if } 1/2 \leq x < 3/4; \\ x - \frac{1}{4}, & \text{if } 3/4 \leq x < 7/8; \\ \vdots & \end{cases}$$

The inductive process should be clear from the definition of T and its partial graph in Figure 3.10: the unit interval is subdivided into a countable number of abutting intervals, starting with $[0, 1/2)$, so that the i^{th} interval has length $1/2^i$. Then T sends $[0, 1/2)$ to $[1/2, 1)$ by the translation map $x \rightarrow x + 1/2$, and the remaining intervals are sent by the corresponding translation to the interval of the same length preceding the previous interval. So, for example, the interval $[1/2, 3/4)$ is sent to the interval $[1/4, 1/2)$ (the interval of length $1/4$ preceding $[1/2, 1)$).

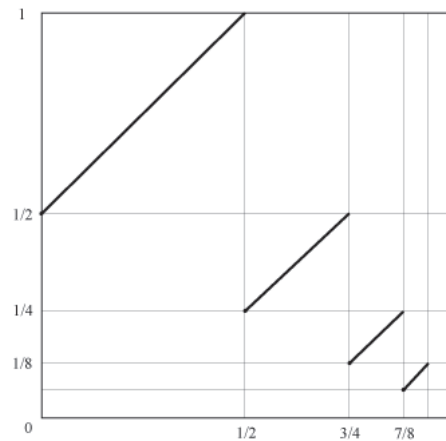


Figure 3.10. Graph of the odometer map

Now we introduce another description of this map. First we need some notation. Given two intervals $I = [a, b)$ and $J = [c, d)$ of the same length, there is a unique orientation-preserving translation from I to J , namely the map $T_{I,J}$ defined by $T_{I,J}(x) = x + c - a$, sending the left point of I to the left point of J . The main properties that we use of $T_{I,J}$ are:

- (1) $T_{I,J}$ is determined by I and J and is a one-to-one map from I onto J ;
- (2) for any measurable set $A \subset I$, $T_{I,J}(A) \subset J$ is measurable and $\lambda(T_{I,J}(A)) = \lambda(A)$ (as Lebesgue measure is translation invariant);

- (3) if I' and J' are dyadic subintervals of I and J , respectively, of the same order, then $T_{I,J}$ agrees with $T_{I',J'}$ on I' .

In the diagrams, the action of $T_{I,J}$ on I will be denoted by an arrow as in Figure 3.11.

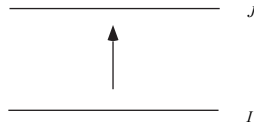


Figure 3.11. Map $T_{I,J}$

With this notation we can describe the value of T on $[0, 1/2)$ by giving two intervals, the intervals $[0, 1/2)$ and $[1/2, 1)$, with the interval $[1/2, 1)$ placed above the interval $[0, 1/2)$ as in Figure 3.12.



Figure 3.12. The odometer on $[0, 1/2)$

To completely define the transformation T we specify a process that generates a sequence of *columns*. A **column**, sometimes also called a **tower**, consists of a finite sequence of disjoint intervals of the same length. The intervals in a column serve to specify the value of the transformation on each interval except the top. Each interval in a column is called a **level**. The number of levels in a column is called the **height** of the column.

The first column will consist of the unit interval, so

$$C_0 = \{[0, 1)\}.$$

(All our intervals will be left-closed and right-open.) We describe how to obtain the next column. It is useful to think of this in terms of cutting and stacking the levels of the previous column. To obtain column C_1 , cut, i.e., divide, the single level in C_0 into the two disjoint subintervals $[0, 1/2)$ and $[1/2, 1)$ and stack them so that the right

subinterval is placed above the left subinterval to form column $C_1 = \{[0, 1/2), [1/2, 1)\}$ of height $h_1 = 2$. Denote the levels of C_1 , from bottom to top, by $I_{1,0}$ and $I_{1,1}$. Column C_1 defines a partial map T_{C_1} by the translation that sends $[0, 1/2)$ to $[1/2, 1)$. Note that T_{C_1} is not defined on $[1/2, 1)$. Figure 3.13 illustrates the process of “cutting and stacking” C_0 to obtain C_1 and also shows the resulting column C_1 .

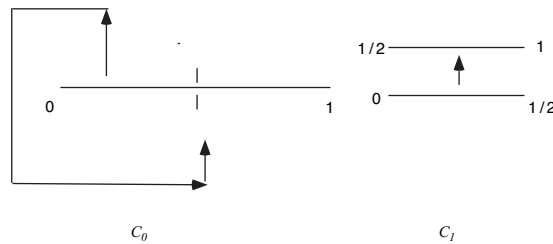


Figure 3.13. Generating column C_1

Before proceeding to the inductive step, to better clarify the construction, we show how to obtain C_2 from C_1 . First, cut each level of C_1 in half, and then stack the right subcolumn above the left subcolumn to obtain the new levels $I_{2,0} = [0, 1/4)$, $I_{2,1} = [1/2, 3/4)$, $I_{2,2} = [1/4, 1/2)$ and $I_{2,3} = [3/4, 1)$. This gives the $h_2 = 4$ levels of C_2 . Observe that T_{C_2} agrees with T_{C_1} wherever T_{C_1} is defined, but T_{C_2} is now defined on the left half of the top level of C_1 ; of course, T_{C_2} remains undefined on the top level of C_2 . Figure 3.14 shows the process of obtaining C_2 from C_1 , and also shows the resulting column C_2 . As one can read column C_2 from the left part of Figure 3.14, this is the figure we find more useful.

Finally, to obtain column C_{n+1} from $C_n = \{I_{n,0}, \dots, I_{n,h_n-1}\}$, cut each level of C_n in half and stack the right subintervals above the left subintervals to obtain a column of height $h_{n+1} = 2h_n$. This is shown in Figure 3.15. One can verify that $h_n = 2^n$.

We have now defined a sequence of partial maps $\{T_{C_n}\}_{n \geq 0}$ so that $T_{C_{n+1}}$ agrees with T_{C_n} wherever T_{C_n} is defined and T_{C_n} is undefined on a subinterval of $[0, 1)$ of measure $\frac{1}{2^n}$ (namely, the top level of C_n). Define the dyadic odometer T by $T(x) = \lim_{n \rightarrow \infty} T_{C_n}(x)$. For each $x \in [0, 1)$ there is some integer $n > 0$ so that x belongs to some level

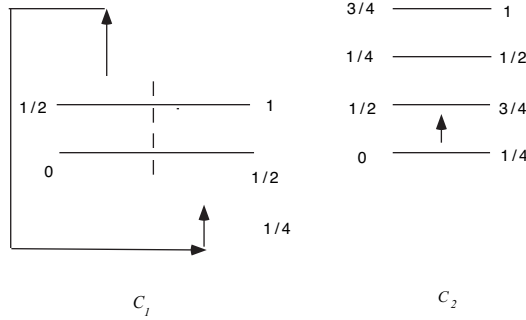


Figure 3.14. Generating column C_2

of C_n that is not the top level. This means that $T(x)$ is well defined. A similar argument shows that T is a one-to-one transformation of $X = [0, 1)$. Furthermore, $T^{-1}(x)$ is defined for all $x \in (0, 1)$. Note that $T^{-1}(0)$ is not defined, but it is not hard to see that T is invertible mod λ . In fact, define X_0 by $X_0 = X \setminus \bigcup_{n=0}^{\infty} \{T^n(0)\}$; in other words, delete from X the endpoints of all the dyadic intervals. Then, $\lambda(X_0) = 1$ and now $T : X_0 \rightarrow X_0$ is one-to-one and onto.

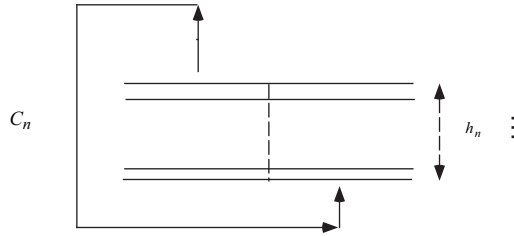


Figure 3.15. Column C_n for the dyadic odometer

Let $n > 0$ and let I be a level in C_n . We introduce some notation to label subintervals in I . When constructing C_{n+1} from C_n , each level in C_n is subdivided into two left-closed right-open, disjoint subintervals that are levels in C_{n+1} ; call the first (leftmost) one $I[0]$ and the second one $I[1]$. So, $I_{n,i}[0]$ is $I_{n+1,i}$ and $I_{n,i}[1]$ is $I_{n+1,i+h_n}$. This notation is extended inductively in the following way. If $I[a_0 a_1 \cdots a_{k-1}]$

(a level in C_{n+k}) has been defined, let $I[a_0a_1 \cdots a_{k-1}0]$ be the left sublevel of $I[a_0a_1 \cdots a_{k-1}]$ in C_{n+k+1} , and similarly let $I[a_0a_1 \cdots a_{k-1}1]$ be the right sublevel of $I[a_0a_1 \cdots a_{k-1}]$ in C_{n+k+1} .

Lemma 3.8.1. *Let T be the dyadic odometer. Then,*

- (1) for all $n > 0$, $T(I_{n,h_n-1}) = I_{n,0}$,
- (2) for all $n > 0$, $i = 0, \dots, h_n - 1$,

$$T^{h_k}(I_{n,i}) = I_{n,i}, \text{ for all } k \geq n.$$

Proof. Let $I = I_{n,h_n-1}$ and $J = I_{n,0}$ be the top and bottom levels of C_n , respectively. From the definition of T_{C_n} , $T(I[0]) = J[1]$. Since this is true for all $n > 0$, using that $I[1], J[0]$ are the top and bottom, respectively, levels of C_{n+1} one obtains that $T(I[1,0]) = T((I[1])[0]) = J[0][1] = J[01]$. Similarly, $T(I[110]) = J[001]$. If we let 1^k denote k consecutive 1's, and similarly for 0^k , by induction we see that $T(I[1^k0]) = J[0^k1]$. As $I[1] = \bigsqcup_{k>0} I[1^k0]$, this completes the proof of part (1). The second part is proved similarly. \square

A consequence of Lemma 3.8.1 is that for each $k > 0$ the transformation T^{h_k} is not ergodic.

The proof of the following lemma is a direct application of the definition and left to the reader. A more general version will be proved later (Lemma 6.5.4).

Lemma 3.8.2. *Let A be a set of positive measure and let I be a dyadic interval that is 3/4-full of A . Let I_0 and I_1 be the first and second half subintervals of I , respectively. Then one of I_0 or I_1 are 3/4-full of A and both I_0 and I_1 are 1/2-full of A .*

Theorem 3.8.3. *If T is the dyadic odometer, then T is a measure-preserving and ergodic invertible mod λ transformation of $[0,1]$.*

Proof. Since the dyadic intervals form a sufficient semi-ring and T is measurable and measure-preserving on the dyadic intervals, Theorem 3.4.1 implies that T is measurable and measure-preserving. That T is invertible mod λ follows from the existence of the set X_0 mentioned above.

To show that T is ergodic let A_1 and B_1 be two sets of positive measure in $[0, 1)$. There exist dyadic intervals I and J such that I and J are $\frac{3}{4}$ -full of A_1 and B_1 , respectively. As before, we may assume that I and J are of the same measure. Thus they are both levels in some column C_{n-1} , say, for some $n > 0$. As each half of J in column C_n is at least $\frac{1}{2}$ -full of B_1 , by considering the appropriate subintervals in column C_n we may finally assume that we have intervals I and J in C_n , each $\frac{1}{2}$ -full of A_1 and B_1 , respectively, and with J above I . Let $A = A_1 \cap I$ and $B = B_1 \cap J$. There is an integer $n > 0$ such that

$$T^n(I) = J.$$

Therefore,

$$\begin{aligned} \lambda(T^n(A_1) \cap B_1) &\geq \lambda(T^n(A) \cap B) \\ &\geq \lambda(T^n(I) \cap J) - \lambda(I \setminus A) - \lambda(J \setminus B) \\ &> \lambda(J) - \frac{1}{2}\lambda(I) - \frac{1}{2}\lambda(J) = 0. \end{aligned}$$

Thus the transformation is ergodic. \square

The construction of the dyadic odometer can be generalized in a natural way. Let $\{r_n\}_{n \geq 0}$ be a sequence of integers ≥ 2 . We let $\{r_n\}_{n \geq 0}$ determine a sequence of columns $\{C_n\}_{n \geq 0}$ in the following way. Let $C_0 = \{[0, 1)\}$. Assuming that C_n has been defined, let C_{n+1} be the column obtained from C_n by cutting each level of C_n into r_n equal-length subintervals and stacking from left to right. This defines a new transformation T called the r_n -**odometer** with $h_{n+1} = r_n h_n$. For example, the dyadic odometer is obtained when $r_n = 2$ for all $n \geq 0$. The exercises explore properties of these transformations.

Exercises

- (1) Give another proof that the dyadic odometer is ergodic by showing directly that if A is a T -invariant set of positive measure, then $\lambda(A) = 1$.
- (2) Let T be the dyadic odometer. Show that for every $n > 0$ and every level I in column C_n , $T^{h_n}(I) = I$. Use this to deduce that for every set of positive measure A and every

integer $k > 0$ there is an integer $n > 0$ such that

$$\lambda(A \cap T^n(A) \cap T^{2n}(A) \cap \dots \cap T^{kn}(A)) > 0.$$

(This is Furstenberg's Multiple Recurrence property for the dyadic odometer.)

- (3) Let T be the dyadic odometer. Show that for all n , the transformation T^{2n} is not ergodic. What about ergodicity of T^k for k odd?
- (4) Construct a finite measure-preserving transformation T so that T and T^2 are ergodic but T^3 is not ergodic.
- (5) Let T be the r_n -adic odometer with $r_n = n$ for $n \in \mathbb{N}$ (i.e., for each $n > 2$ column C_n is cut into $r_n = n$ subcolumns). Show that the transformation T^k is ergodic if and only if $k = 1$.
- (6) Let T be the shift on \mathbb{Z} with counting measure. Is T totally ergodic?
- (7) Let T be the dyadic odometer. For each $n > 0$, extend the column map T_{C_n} so that it is defined on the top level of C_n by the translation that takes the interval I_{n,h_n-1} to the interval $I_{n,0}$. (Note that while T also takes I_{n,h_n-1} to $I_{n,0}$, it is not a translation on I_{n,h_n-1} and differs from T_{C_n} on points of this interval.) a) Show that the extended map T_{C_n} is a nonergodic finite measure-preserving map of $[0, 1)$. b) Show that the sequence of maps T_{C_n} converges to T in the sense that for every measurable set A ,

$$\lim_{n \rightarrow \infty} \lambda(T_{C_n}(A) \triangle T(A)) = 0.$$

3.9. Infinite Measure-Preserving Transformations

We have seen that an infinite measure-preserving transformation does not have to be recurrent. For example, $T(x) = x + 1$ on \mathbb{R} with Lebesgue measure is measure-preserving but not recurrent. Also, $T(n) = n + 1$ on \mathbb{Z} with counting measure is measure-preserving and ergodic but not recurrent. We start with a lemma showing that for invertible transformations on nonatomic spaces, ergodicity implies

recurrence. We then present an example of an invertible measure-preserving transformation on the positive reals that is ergodic.

Lemma 3.9.1. *Let T be an invertible measure-preserving transformation on a nonatomic σ -finite measure space (X, \mathcal{S}, μ) . If T is ergodic, then it is recurrent.*

Proof. Suppose that T is not recurrent. Then there exists a set A of positive measure such that $\mu(T^{-n}(A) \cap A) = 0$ for all $n > 0$. It follows that $\mu(\bigcup_{n \neq 0} (T^n(A) \cap A)) = 0$ (here the notation means that the union is taken over all positive and negative n that are nonzero). Then the set

$$W = A \setminus \left(\bigcup_{n \neq 0} (T^n(A) \cap A) \right)$$

satisfies

$$\begin{aligned} \mu(W) &= \mu(A) > 0 \text{ and} \\ T^n(W) \cap T^m(W) &= \emptyset \text{ for all } m \neq n. \end{aligned}$$

Since (X, \mathcal{S}, μ) is nonatomic, there exists $B \subset W$ such that $0 < \mu(B) < \mu(W)$. The set $B^* = \bigcup_{n=-\infty}^{\infty} T^n(B)$ is T -invariant and one can verify that $\mu(B^*) > 0$ and $\mu((B^*)^c) > 0$. This contradicts the fact that T is ergodic. Therefore T is recurrent. \square

We construct now an ergodic infinite measure-preserving invertible transformation T that was introduced by Hajian and Kakutani in 1970. The construction of this transformation uses the method of cutting and stacking; we will need to inductively define a sequence of columns $\{C_n\}$. While in the case of the dyadic odometer the union of the levels of any column is always the interval $[0, 1)$, here column C_{n+1} will contain new levels that are not contained in the union of the levels of C_n ; these new levels are called **spacers**. The union of the levels over all the columns will be $[0, \infty)$. Thus the transformation will be defined on an infinite measure space.

We start by letting $C_0 = \{[0, 1)\}$, $h_0 = 1$ and $X_0 = [0, 1)$. Before giving the inductive step, we explain how column C_1 is obtained from C_0 . As in the case of the dyadic odometer, subdivide $[0, 1)$ into the equal length subintervals $[0, 1/2)$, $[1/2, 1)$, but in this case

we add two new subintervals of length $1/2$. We choose the subintervals abutting $[0, 1)$ so that they are $[1, 3/2)$ and $[3/2, 2)$. Then $C_1 = \{[0, 1/2), [1/2, 1), [1, 3/2), [3/2, 1)\}$; the height is $h_1 = 4h_0 = 4$.

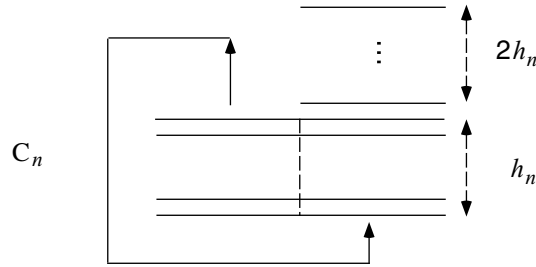


Figure 3.16. Generating C_{n+1} from C_n in the Hajian–Kakutani skyscraper

Given a column C_n with height h_n , write $C_n = \{I_{n,0}, \dots, I_{n,h_n-1}\}$, where $X_n = \bigsqcup_{j=0}^{h_n-1} I_{n,j}$ is an interval. To construct C_{n+1} first cut each level $I_{n,j}$ of C_n into two equal-length subintervals $I_{n,j}^{[0]}$ and $I_{n,j}^{[1]}$. Then choose $2h_n$ new intervals abutting the right endpoint of X_n , each of the same length as any of the intervals $I_{n,j}^{[i]}$, and denote these spacers by $K_{1,0}, \dots, K_{1,2h_n-1}$ (we will see later why we choose the first subscript to be 1). The levels of C_{n+1} are then

$$\begin{aligned} I_{n+1,0} &= I_{n,0}^{[0]}, \dots, I_{n+1,h_n-1} = I_{n,h_n-1}^{[0]}, \\ I_{n+1,h_n} &= I_{n,0}^{[1]}, \dots, I_{n+1,2h_n-1} = I_{n,h_n-1}^{[1]}, \\ I_{n+1,2h_n} &= K_{1,0}, \dots, I_{n+1,4h_n-1} = K_{1,2h_n-1}. \end{aligned}$$

Then $T_{C_{n+1}}$ is defined as the translation that sends each level in C_{n+1} to the level above it. It is clear that $T_{C_{n+1}}$ agrees with T_{C_n} wherever this map was defined and $T_{C_{n+1}}$ is now defined on $I_{n,2h_n}^{[0]}, I_{n,2h_n}^{[1]}$, and all the new spacer levels except the top one. Write $X_{n+1} = \bigsqcup_{j=0}^{4h_n-1} I_{n+1,j}$; it is left to the reader to verify that X_{n+1} is an interval and $\lambda(X_{n+1}) = 2\lambda(C_n) = 2^{n+1}$. It follows that the transformation $T(x) = \lim_{n \rightarrow \infty} T_{C_n}(x)$ is defined on $X = \bigcup_{n=0}^{\infty} X_n = [0, \infty)$.

There is another picture that is useful to keep in mind for this transformation. We think of a “tower” (or column) defined over the unit interval in the following way. Over the subinterval $[1/2, 1)$ place

two intervals of length $1/2$. These intervals are chosen to be $[1, 3/2)$ and $[3/2, 2)$. We think of the transformation as moving points up as long as there is an interval to go to. Next, above the interval $[7/4, 2)$ place 8 intervals of length $1/4$.

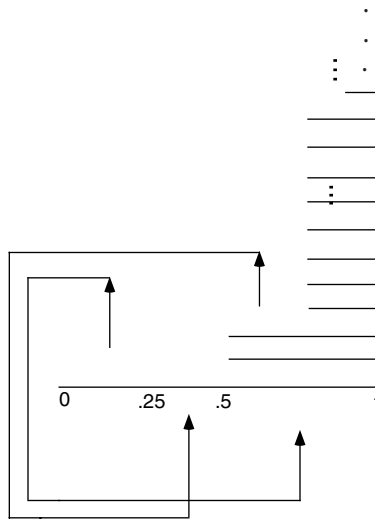


Figure 3.17. The Hajian-Kakutani transformation

A set W is said to be **weakly wandering** under an invertible transformation T if there is a sequence $\{n_i\}_{i=1}^{\infty} \geq 0$ such that

$$T^{n_i}(W) \cap T^{n_j}(W) = \emptyset, \text{ for } n_i \neq n_j.$$

The weakly wandering set W is said to be **exhaustive** if for the weakly wandering sequence $\{n_i\}$

$$X = \bigcup_{i=1}^{\infty} T^{n_i}(W) \text{ mod } \lambda.$$

The reader is asked in the exercises to verify that a finite measure-preserving transformation does not admit weakly wandering sets. A related notion is that of a wandering set. A set W is called **wandering** if $T^{-n}(W) \cap T^{-m}(W) = \emptyset$ for all $m \neq n$, for $n, m \in \mathbb{N}$. Evidently, a wandering set is weakly wandering. In Exercise 6 the

reader is asked to show that a recurrent transformation does not admit a wandering set of positive measure. If a transformation is invertible, a wandering set also satisfies: $T^{-n}(W) \cap T^{-m}(W) = \emptyset$ for all $m \neq n$. It can be shown, as an application of the ergodic theorem for infinite measure-preserving transformations, that every ergodic infinite measure-preserving invertible transformation admits a weakly wandering set of positive measure. We prove below the existence of weakly wandering sets for the case of the Hajian–Kakutani transformation.

Theorem 3.9.2. *The Hajian–Kakutani transformation T is infinite measure-preserving, invertible mod λ , recurrent and ergodic. Furthermore, T admits an exhaustive weakly wandering set of measure 1.*

Proof. Since the levels in all columns form a sufficient semi-ring, and T sends levels to levels of the same measure, Theorem 3.4.1 implies that T is measure-preserving. It is also clear that T is invertible mod λ . The proof of ergodicity is similar to the one for the dyadic odometer. Let A, B be sets of positive measure. There exists a column C_n and levels I and J in C_n such that I is above J and they are both $1/2$ -full of A and B , respectively. There is an integer $k > 0$ such that $T^k(J) = I$. Then $\lambda(T^k(A) \cap B) > 0$.

We shall now show that $W = [0, 1)$ is an exhaustive weakly wandering set. Note that the number of spacers that are added to column i to form column $i+1$ is $s_i = 2^{2^{i+1}}$; i.e., there are twice as many as the number of levels of column i . It follows that the sequence of the number of spacers is $2^1, 2^3, 2^5, \dots$. We shall see that the weakly wandering sequence is an appropriate sum of elements of this sequence.

We construct a sequence of positive integers $\{n_i\}_{i \geq 0}$ such that

$$T^{n_i}(W) \cap T^{n_j}(W) = \emptyset \text{ for } i \neq j.$$

Recall that any integer $i \geq 0$ has a unique base-2 representation of the form

$$i = \delta_0 \cdot 2^0 + \delta_1 \cdot 2^1 + \dots + \delta_k \cdot 2^k,$$

where $\delta_j \in \{0, 1\} (j = 1, \dots, k)$ and k is some integer depending on i . Then define

$$n_i = \delta_0 \cdot 2^1 + \delta_1 \cdot 2^3 + \dots + \delta_k \cdot 2^{2k+1}.$$

The proof now follows by verifying inductively on n that the sets $T^{n_i}(W), i = 1, \dots, 2^n - 1$, are disjoint, and that $\bigsqcup_{i=1}^{2^n-1} T^{n_i}(W)$ is equal to the union of the levels in C_n . \square

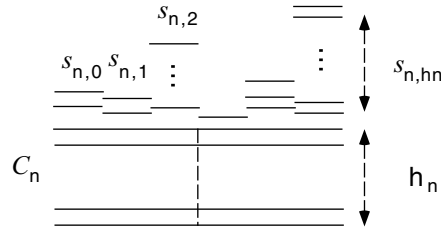


Figure 3.18. Cutting and stacking the C_n column

We now describe a more general type of construction. We again start with C_0 consisting of an interval, usually $[0, 1)$. Given a column C_n , to obtain column C_{n+1} , in addition to the number of cuts r_n , we need to specify the number of spacers that are added on top of each subcolumn. So let $\{s_{n,i}\}$ be a doubly-indexed sequence of nonnegative integers for $n \geq 0$ and $i = 0, \dots, r_n - 1$. First cut each level of C_n into r_n subintervals to obtain r_n subcolumns of C_n indexed from 0 to $r_n - 1$. Then add on top of the i^{th} -subcolumn $s_{n,i}$ spacers (of the same length as all the levels in that subcolumn). Then extend the transformation to send, by translation, the top level of the i^{th} -subcolumn to the first spacer above, and its top spacer to the bottom level of the $i + 1^{\text{st}}$ -subcolumn. The map $T_{C_{n+1}}$ remains undefined on the top level of the last subcolumn if there are no spacers above it, and on the top spacer of the last subcolumn if there are spacers about the last subcolumn. Figure 3.18 illustrates this. In particular we note that the Hajian–Kakutani skyscraper is obtained for $r_n = 2$ and $s_{n,0} = 0, s_{n,1} = 2h_n$. We will call this construction the **cutting and stacking** construction with sequence of cuts r_n and sequence

of **spacers** $s_{n,i}$. In Chapter 6, we will see constructions where C_0 consists of an interval different from $[0, 1)$.

Exercises

- (1) Let T be the Hajian–Kakutani skyscraper transformation. Show that T^2 is not ergodic.
- (2) Let T be the Hajian–Kakutani skyscraper transformation. Find an exhaustive weakly wandering set of measure 2 for T . Is there an exhaustive weakly wandering set of infinite measure for T ?
- (3) Construct an ergodic infinite measure-preserving transformation T such that T^2 is ergodic but T^3 is not ergodic.
- (4) Let T be a cutting and stacking construction with $r_n = 2$ and $s_{n,0} = 0, s_{n,1} = 2h_n + 1$. Show that T^n is ergodic for all $n \neq 0$.
- (5) Let T be the Hajian–Kakutani skyscraper transformation. Find all integers k so that T^k is ergodic.
- (6) Let T be a measure-preserving transformation. Show that T is recurrent if and only if T does not admit any wandering sets of positive measure.

3.10. Factors and Isomorphism

We study what it means for two dynamical systems to be “the same.” The technical term will be isomorphic. The expectation is that two isomorphic dynamical systems will share the same dynamical properties. For example, if a system is ergodic, its isomorphic systems should be ergodic. To start with a simple example, consider a transformation T defined on a space X consisting of two points $X = \{a, b\}$ and a measure defined by $\mu(\{a\}) = \mu(\{b\}) = \frac{1}{2}$ and such that $T(a) = b$ and $T(b) = a$. Define another transformation $S : Y \rightarrow Y$ on $Y = \{x, y, z\}$ by $S(x) = y, S(y) = x, S(z) = z$ and a measure ν given by $\nu(\{x\}) = \nu(\{y\}) = \frac{1}{2}, \nu(\{z\}) = 0$. It is clear that $T : X \rightarrow X$ and $S : Y \rightarrow Y$ should be considered to be isomorphic dynamical systems. First, note that the measure spaces X, μ and Y, ν , after renaming the points and discarding the null sets, are the same. Next note that the map

$\phi : X \rightarrow Y$ defined by $\phi(a) = x, \phi(b) = y$ preserves the dynamical structure of the maps, in that $T(\phi(a)) = \phi(S(a)), T(\phi(b)) = \phi(S(b))$. Furthermore, the map ϕ is measure-preserving and invertible (after discarding the null set $\{c\}$).

We start with the notion that is used to identify measure spaces. Two measure spaces (X, \mathcal{S}, μ) and (X', \mathcal{S}', μ') are said to be **isomorphic** (sometimes we may say measure-theoretically isomorphic or isomorphic mod 0) if there exist measurable sets $X_0 \subset X$ and $X'_0 \subset X'$ of **full measure** (i.e., $\mu(X \setminus X_0) = 0$ and $\mu'(X' \setminus X'_0) = 0$) and a map

$$\phi : X_0 \rightarrow X'_0$$

that is one-to-one and onto and such that

- (1) $A \in \mathcal{S}'(X'_0)$ if and only if $\phi^{-1}(A) \in \mathcal{S}(X_0)$,
- (2) $\mu(\phi^{-1}(A)) = \mu'(A)$ for all $A \in \mathcal{S}'(X'_0)$.

We call the map ϕ a **measure-preserving isomorphism mod 0** or **isomorphism mod 0**. Sometimes when we want to be specific about the measures we shall write mod (μ, μ') instead of mod 0.

Example. Let $X = [-1, 1], \mathcal{S} = \mathfrak{L}([-1, 1])$, and define a measure μ on \mathcal{S} by $\mu(A) = \frac{1}{2}\lambda(A)$, for $A \in \mathcal{S}$. Clearly, (X, \mathcal{S}, μ) is a measure space; we claim it is isomorphic to $([0, 1], \mathfrak{L}([0, 1]), \lambda)$. Indeed, let $\phi : [0, 1] \rightarrow X$ be given by $\phi(x) = 2x - 1$. Evidently, ϕ is one-to-one and onto and we can compute its inverse $\phi^{-1}(y) = (y + 1)/2$. It is not hard to see that ϕ and ϕ^{-1} are measurable. This can be seen by appealing to Exercise 2.3.1 (a result analogous to Theorem 3.4.1 but whose formulation is left to the reader). That ϕ is measure-preserving follows from the fact that it is the composition of a translation (which leaves Lebesgue measure invariant) and a dilation (which expands the measure by 2). So, $\lambda(\phi^{-1}(A)) = \frac{1}{2} \cdot \lambda(A) = \mu(A)$ for all $A \in \mathcal{S}$. (A result analogous to Theorem 3.4.1 also holds here, i.e., for maps from one measure space to another, stating that it suffices to verify the measure-preserving property on a sufficient semi-ring.)

Example. Let F be a closed set in $[0, 1]$ of positive Lebesgue measure. Define $\mu(A) = \lambda(A)/\lambda(F)$ for $A \in \mathfrak{L}(F)$. Then $(F, \mathfrak{L}(F), \mu)$ is isomorphic to $([0, 1], \mathfrak{L}([0, 1]), \lambda)$. The isomorphism we define is an

interesting map that has other applications. It is a map $\phi : F \rightarrow [0, 1]$ defined by

$$\phi(x) = \frac{\lambda(F \cap [0, x])}{\lambda(F)}.$$

An important property of ϕ is that it is continuous. Continuity follows from the fact that λ is a nonatomic measure. Let $\varepsilon > 0$. If $|x - y| < \varepsilon$, say $x \geq y$, then

$$\begin{aligned} |\phi(x) - \phi(y)| &= \lambda(F \cap [0, x]) - \lambda(F \cap [0, y]) \\ &= \lambda(F \cap [y, x]) \leq |x - y| < \varepsilon. \end{aligned}$$

This proof shows that ϕ is uniformly continuous. Next we note that ϕ is nondecreasing: if $x < y$, $x, y \in F$, then $\mu(F \cap [0, x]) \leq \mu(F \cap [0, y])$. To see that ϕ is onto, we use the fact that ϕ is nondecreasing. Let $\beta = \sup F$; β is a number in F as $F \subset [0, 1]$ is closed. From the definition of β it is clear that $\phi(\beta) = 1$. Similarly, if $\alpha = \inf F$, then $\alpha \in F$ and $\phi(\alpha) = 0$. Therefore ϕ must be onto. The measure-preserving property can be verified on the sufficient semi-ring of intervals in $[0, 1]$. Let $I = (a, b) \subset [0, 1]$ and choose $a_0, b_0 \in F$ so that $\phi(a_0) = a$, $\phi(b_0) = b$. Then, $\phi^{-1}([0, b]) = \{x \in F : 0 \leq \phi(x) < \phi(b_0)\} = F \cap [0, b_0)$. So

$$\begin{aligned} \mu(\phi^{-1}([a, b])) &= \mu(\phi^{-1}[0, b] \setminus \phi^{-1}[0, a]) = \mu(F \cap [0, b_0) \setminus F \cap [0, a_0)) \\ &= \mu(F \cap [0, b_0)) - \mu(F \cap [0, a_0)) = \phi(b_0) - \phi(a_0) \\ &= b - a = \lambda([a, b]). \end{aligned}$$

It is not hard to verify that properties of a measure space such as sigma-finiteness and the number of atoms, for example, are preserved under isomorphism. The property of being complete (for a measure space), however, is not preserved.

Now we are in a position to define the most important type of measure spaces that we consider. A complete measure space (X, \mathcal{S}, μ) is called a **Lebesgue space** if it is isomorphic mod 0 to a canonical Lebesgue measure space. We note here that we allow Lebesgue spaces to be of infinite σ -finite measure, while sometimes in the literature they are assumed to be of finite measure. The most interesting measure spaces are Lebesgue spaces and one can make the case that they are the only spaces one needs to be concerned with in ergodic theory. For example, in the exercises the reader is asked to show that the

unit interval with Lebesgue measure is isomorphic to the unit square with Lebesgue measure, and that \mathbb{R} is isomorphic to \mathbb{R}^d .

A **measure-preserving dynamical system** consists of a Lebesgue measure space (X, \mathcal{S}, μ) and a measure-preserving transformation $T : X \rightarrow X$. In a similar way we define an invertible measure-preserving dynamical system and an invertible measure-preserving dynamical system mod 0.

We now consider the notion of isomorphism for dynamical systems. For the remainder of this section we only consider finite measure-preserving transformations.

Let (X, \mathcal{S}, μ, T) and $(X', \mathcal{S}', \mu', T')$ be two finite measure-preserving dynamical systems. We say that the two systems are **isomorphic** if there exist measurable sets $X_0 \subset X$ and $X'_0 \subset X'$ of full measure (i.e., $\mu(X \setminus X_0) = 0$ and $\mu'(X' \setminus X'_0) = 0$) with

$$T(X_0) \subset X_0, T'(X'_0) \subset X'_0,$$

and there exists a map $\phi : X_0 \rightarrow X'_0$, called an **isomorphism**, that is one-to-one and onto and such that for all $A \in \mathcal{S}'(X'_0)$,

- (1) $\phi^{-1}(A) \in \mathcal{S}(X_0)$,
- (2) $\mu(\phi^{-1}(A)) = \mu'(A)$, and
- (3) $\phi(T(x)) = T'(\phi(x))$ for all $x \in X_0$.

The role of the set X_0 is to make precise the fact that the properties of the isomorphism need to hold only on a set of full measure. Property (3) is called **equivariance** and is illustrated the following diagram (note that one typically writes the sets X and X' though the property only holds for a set of full measure in X and X'):

$$(3.7) \quad \begin{array}{ccc} X & \xrightarrow{T} & X \\ \phi \downarrow & & \downarrow \phi \\ X' & \xrightarrow{T'} & X' \end{array}$$

We shall understand a **dynamical property** to be a property that is invariant under isomorphism. That is, if one dynamical system

exhibits the property, then so do all isomorphic systems. (Of course, we refer here to measurable dynamical properties.)

A related and important concept is that of a *factor*. A factor map is similar to an isomorphism except that the factor map is not required to be one-to-one. A dynamical system $(X', \mathcal{S}', \mu', T')$ is a **factor** of (X, \mathcal{S}, μ, T) if there exist measurable sets $X_0 \subset X$ and $X'_0 \subset X'$ of full measure with

$$T(X_0) \subset X_0, T'(X'_0) \subset X'_0,$$

and a map $\phi : X_0 \rightarrow X'_0$ that is onto and such that for all $A \in \mathcal{S}'(X'_0)$,

$$(1) A \in \mathcal{S}'(X'_0) \text{ if and only if } \phi^{-1}(A) \in \mathcal{S}(X_0),$$

$$(2) \mu(\phi^{-1}(A)) = \mu'(A) \text{ for all } A \in \mathcal{S}'(X'_0).$$

The main difference with an isomorphism is that the factor map is not required to be one-to-one a.e.

Example. Let $Y = \{p\}$ be a one-point space and let ν be a probability measure on Y . Let $S : Y \rightarrow Y$ be the identity transformation. Then $(Y, \mathcal{P}(Y), \nu, S)$ is a factor of any dynamical system (X, \mathcal{S}, μ, T) ; it is called the trivial factor. Evidently, T is also a factor of itself.

Example. Let T be the dyadic odometer. From the definition of T it follows that for each $n > 1$ and $i \in \{0, \dots, h_n - 1\}$, $T(I_{n,i}) = I_{n,i+1}$. This suggests that T acts on the elements of C_n as a rotation. We make this more explicit by defining a map $\phi : X \rightarrow Z_n$ by $\phi(x) = i$ if $x \in I_{n,i}$ (recall $Z_n = \{0, \dots, h_n - 1\}$). We claim that ϕ is a factor map and, in this way, see the rotation on n points as a factor of T .

Lemma 3.10.1. *Let S be a factor of T . Then if T is ergodic so is S .*

Proof. Let $T : (X, \mu) \rightarrow (Y, \nu)$ and $\phi : X \rightarrow Y$ be the factor map. If A is a strictly invariant set for S , then we claim that $\phi^{-1}(A)$ is a strictly invariant set for T . In fact, $T^{-1}(\phi^{-1}(A)) = \phi^{-1}(S^{-1}(A)) = \phi^{-1}(A)$. Also, $\mu(\phi^{-1}(A)) = \nu(A)$. So $\nu(A) = 0$ or $\nu(A^c) = 0$. \square

Exercises

- (1) Show that if (X, \mathcal{S}, μ, T) is a Lebesgue measure-preserving dynamical system, then for any $X_0 \in \mathcal{S}(X)$ with $T^{-1}(X_0) = X_0$, the system $(X_0, \mathcal{S}(X_0), \mu, T)$ is a measure-preserving dynamical system.
- (2) Let (X, \mathcal{S}, μ) be a Lebesgue measure space and let $X_0 \in \mathcal{S}(X)$ with $\mu(X \setminus X_0) = 0$. Suppose there exists a transformation T_0 so that $(X_0, \mathcal{S}(X_0), \mu, T_0)$ is a measure-preserving dynamical system. Show that there exists a transformation $T : X \rightarrow X$ so that $T(x) = T_0(x)$ for $x \in X_0$ and (X, \mathcal{S}, μ, T) is a measure-preserving dynamical system. (T is not unique but differs from T_0 on only a null set.) Show that T_0 is isomorphic to T .
- (3) Show that if T and S are isomorphic, then so are T^n and S^n for all $n > 0$.
- (4) Show that the doubling map is a factor of the baker's transformation.
- (5) Is the dyadic odometer isomorphic to an irrational rotation?
- (6) Let T be an ergodic finite measure-preserving transformation. Show that if T^i is ergodic for $i = 1, \dots, k-1$ but T^k is not ergodic, then there exists a measurable set A such that the set $T^i(A), i = 1, \dots, k-1$, are disjoint mod μ and $\bigcup_{i=0}^{k-1} T^i(A) = X \text{ mod } \mu$.
- * (7) Let T be an ergodic finite measure-preserving transformation. Show that T is totally ergodic if and only if it has no factor that is a rotation on n points for any $n > 1$.

3.11. The Induced Transformation

This section treats induced transformations, a useful construction introduced by Kakutani in 1941. While we state the definition in the general case, for most of the section we restrict ourselves to invertible transformations on finite measure spaces, as these illustrate the main ideas and are significantly simpler.

Let (X, \mathcal{S}, μ) be a σ -finite measure space and let $T : X \rightarrow X$ be a recurrent measure-preserving transformation. Then, for every measurable set A of positive measure there is a null set $N \subset A$ such that for all $x \in A \setminus N$ there is an integer $n = n(x) > 0$ with $T^n(x) \in A$. We call the smallest such n the **first return time** to A , defined by

$$n_A(x) = \min\{n > 0 : T^n(x) \in A\}.$$

We think of $n_A(x)$ as the first “time” that x comes back to A under iteration by T . Since T is recurrent, $n_A(x)$ is defined a.e. for all sets A of positive measure.

As $n_A(x)$ is defined a.e. for all sets A of positive measure, it is possible to define the **induced transformation** of T on A . The induced transformation on a set A of positive measure is denoted by T_A and is defined by

$$T_A(x) = T^{n_A(x)}(x) \text{ for a.e. } x \in A.$$

To define T_A on every point of A one can just let $T_A(x)$ have any fixed value for all x in the null set where n_A is not defined.

Proposition 3.11.1. *Let (X, \mathcal{S}, μ) be a finite measure space and $T : X \rightarrow X$ an invertible measure-preserving transformation. If T is a recurrent transformation and A a set of positive measure, then the induced transformation T_A is measure-preserving on A .*

Proof. For each integer $i \geq 1$ let

$$A_i = \{x \in A : n_A(x) = i\}.$$

The sets A_i are disjoint and

$$A = \bigsqcup_{i=1}^{\infty} A_i \text{ mod } \mu.$$

Let $B \subset A$ be measurable. Then,

$$\begin{aligned} \mu(T_A(B)) &= \mu(T_A(\bigsqcup_{n=1}^{\infty} B \cap A_n)) = \mu(\bigsqcup_{i=1}^{\infty} T_A(B \cap A_i)) \\ &= \sum_{i=1}^{\infty} \mu(T^i(B \cap A_i)) = \sum_{i=1}^{\infty} \mu(B \cap A_i) \\ &= \mu(\bigsqcup_{i=1}^{\infty} \mu(B \cap A_i)) = \mu(B). \end{aligned}$$

□

When T_A is defined on a set of finite positive measure A , one usually defines a normalized probability measure μ_A on A by $\mu_A(B) = \mu(B)/\mu(A)$, and T_A is considered with the measure μ_A .

Let T be an invertible, recurrent, measure-preserving transformation. If T is ergodic and A is any set of positive measure, then $\bigcup_{i=1}^{\infty} T^i(A) = X \bmod \mu$. As before let $A_i = \{x \in A : n_A(x) = i, \text{ for } i \geq 1\}$. Then $A = \bigsqcup_{i=1}^{\infty} A_i \bmod \mu$. Therefore, up to a set of μ -measure 0, the set X has the structure given by Figure 3.19. We can observe that if a point x is in A_i , the sets $T(A_i), \dots, T^{i-1}(A_i)$ are disjoint and outside A . We think of this as a column above i . Once a point reaches the top of the column, its next iterate brings it down to A . Using this, one can obtain a simple proof of the following lemma.

Lemma 3.11.2. *Let (X, \mathcal{S}, μ, T) be an invertible, finite measure-preserving dynamical system. If T is an ergodic transformation and A a set of positive measure, then the transformation T_A is ergodic on A .*

Proof. Let E, F be sets of positive measure in A . As T is ergodic (and recurrent) there exists an integer $n > 0$ such that $\mu(T^n(E) \cap F) > 0$. So there is a point $x \in E$ such that $T^n(x) \in F$. Let $n_1 = n_A(x), n_2 = n_A(T^{n_1}(x)), \dots$ until n_k (is the first integer) so that $T^{n_k}(x) \in F$. Then $T_A^k(x) = T^n(x) \in F$, showing that $T_A^k(F) \cap E \neq \emptyset$. Therefore T_A is ergodic. □

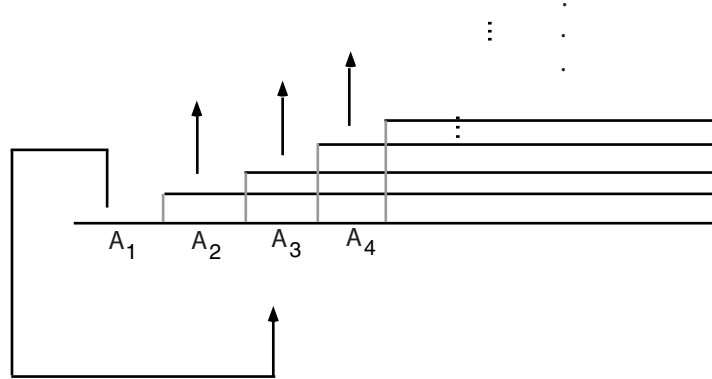


Figure 3.19. An ergodic transformation over a set A

Exercises

- (1) Show that every set of positive Lebesgue measure in \mathbb{R} contains a non-Lebesgue measurable subset.
- (2) Let T be the dyadic odometer and $A = [0, \frac{1}{2})$. Show that T is isomorphic to the induced transformation T_A on A with normalized measure $\mu_A(B) = \mu(B)/\mu(A)$.
- (3) Let (X, \mathcal{S}, μ, T) be an invertible, recurrent, finite measure-preserving dynamical system. If A is a set of positive measure such that the transformation T_A is ergodic on A , then show that T is ergodic.

3.12. Symbolic Spaces

We study some important symbolic spaces, which at the same time provide interesting examples of metric spaces. Let $N > 1$ be an integer. We let Σ_N^+ denote the set consisting of all infinite sequences of symbols from $\{0, \dots, N-1\}$. Each element x of Σ_N^+ has the form $x = x(0)x(1)\cdots$ where $x(i) \in \{0, \dots, N-1\}$ for $i \geq 0$. Sometimes it helps to think of x as a function $x : \mathbb{N}_0 \rightarrow \{0, \dots, N-1\}$, where we write $x(i)$ for the value of x at $i \in \mathbb{N}$. We can also view Σ_N^+ as a

countably infinite Cartesian product of the finite set $\{0, \dots, N-1\}$, but we do not emphasize this definition.

While we will concentrate on the space of one-sided infinite sequences Σ_N^+ , the space of two-sided infinite sequences, denoted by Σ_N , is also very important. We define the space Σ_N of two-sided infinite sequences of symbols from $\{0, \dots, N-1\}$ in a similar way: each element of Σ_N can be seen as a function $x: \mathbb{Z} \rightarrow \{0, \dots, N-1\}$, and we write the element x as

$$\cdots x(-2)x(-1)x(0)x(1)x(2)\cdots$$

The properties of this space are similar to those of Σ_N^+ and are left to the reader as exercises.

We describe a metric in Σ_N^+ . Given $x, y \in \Sigma_N^+$, let

$$I(x, y) = \min\{i \geq 0 : x(i) \neq y(i)\}.$$

We think of $I(x, y)$ as the first time where x and y differ. Then define a metric d on Σ_N^+ by

$$d(x, y) = 2^{-I(x, y)},$$

if $x \neq y$ and $d(x, x) = 0$.

Example. Let $x = 0101\overline{01}\cdots$, $y = 01011\overline{1}\cdots$, $z = 10101\overline{1}\cdots$, where \overline{a} means that the pattern a is repeated. Then

$$d(x, y) = 2^{-4}, d(x, z) = 2^0 = d(y, z).$$

The ball $B(x, 1)$ consists of all elements of Σ_N^+ starting with 0. Note also that $B(x, 1) = B(x, 3/4)$ and $B(x, 3/2) = \Sigma_N^+$.

Lemma 3.12.1. (Σ_N^+, d) is a compact metric space.

Proof. It is clear that $d(x, y) = 0$ if and only if $x = y$, and $d(x, y) = d(y, x)$. To show the triangle inequality, note that $I(x, y)$ is the first integer where x and y differ (assuming $x \neq y$). Let $I(x, z) = i$ and $I(z, y) = j$. If $j \geq i$, then y and z cannot differ before i , or $I(x, y) \geq i$. Thus, $I(x, y) \geq \min\{I(x, z), I(y, z)\}$. If $i > j$, then $I(x, y) \geq j$. So in this case also, $I(x, y) \geq \min\{I(x, z), I(y, z)\}$. It follows that

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}.$$

This inequality implies the triangle inequality. So d is a metric. (A metric satisfying this stronger inequality is called an *ultrametric*.)

To show compactness, let $\{x_n\}$ be a sequence of elements of Σ_N^+ . The idea of the argument can be described in a simpler way by assuming $N = 2$. In the sequence $\{x_n(0)\}$ of first entries there must be infinitely many n so that $x_n(0)$ consists of 0 or of 1 (or both). Suppose 0 is the symbol that appears infinitely often. Let x_{n_1} be the first element of the sequence whose first entry is 0. Now consider the infinite subsequence of those elements that start with 0. Of these, infinitely many must start with 00 or with 01. Say it is 01. Let x_{n_2} be the first element of this new subsequence starting with 01. Of the infinitely many elements starting with 01, there must be infinitely many starting with 010 or 011. In this way we choose the next element of the sequence. Continue in this way to construct the sequence x_{n_i} . Let x be the element of Σ_2^+ such that $x(i) = x_{n_i}(i)$. It is clear that x_{n_i} converges to x in Σ_2^+ . The proof for Σ_N^+ is analogous. \square

We observe that a metric space that is compact must be complete, as any Cauchy sequence must have a convergent subsequence to a point x of the space. But then one can show that the whole Cauchy sequence must converge to x . The following lemma further explores the topology of Σ_N^+ .

Lemma 3.12.2. *Let $B(x, \varepsilon)$ be an open ball in Σ_N^+ . Then $B(x, \varepsilon)$ is a closed set.*

Proof. Let $k > 0$ be such that $1/2^k < \varepsilon \leq 1/2^{k-1}$. Then $y \in B(x, \varepsilon)$ if and only if x and y agree at least in the entries $0, \dots, k$. If x_n is a sequence in $B(x, \varepsilon)$ converging to z , then z must also agree with x in the entries $0, \dots, k$, so it must be an element of $B(x, \varepsilon)$. \square

We present a useful description of balls. Let $w = a_0 \cdots a_{k-1}$ be a word consisting of $k > 0$ symbols from $\{0, \dots, N-1\}$. The **cylinder** based at w is

$$[a_0 \cdots a_{k-1}] = \{x \in \Sigma_N^+ : x_0 = a_0, \dots, x_{k-1} = a_{k-1}\}.$$

So, for example, $\Sigma_2^+ = [0] \sqcup [1]$, and $[1] = [10] \sqcup [11]$.

Lemma 3.12.3 gives a useful characterization of the notion of continuity in metric spaces.

Lemma 3.12.3. *Let (X, d) and (Y, q) be two metric spaces and $\phi : X \rightarrow Y$. The map ϕ is continuous on X if and only if for all $y \in Y$, $\varepsilon > 0$, the set $\phi^{-1}(B(y, \varepsilon))$ is open in X .*

Proof. Suppose that ϕ is continuous and let $y \in Y$, $\varepsilon > 0$. If $\phi^{-1}(B(y, \varepsilon)) = \emptyset$, then it is open. Otherwise let z be a point in $\phi^{-1}(B(y, \varepsilon))$; so $q(\phi(z), y) < \varepsilon$. Since ϕ is continuous, as $\varepsilon - q(\phi(z), y) > 0$, there exists $\delta > 0$ so that if $d(z, x) < \delta$, then $q(\phi(z), \phi(x)) < \varepsilon - q(\phi(z), y)$. So if $x \in B(z, \delta)$, then $q(\phi(x), \phi(z)) < \varepsilon - q(\phi(z), y)$. Thus, $q(\phi(x), y) \leq q(\phi(x), \phi(z)) + q(\phi(z), y) < \varepsilon - q(\phi(z), y) + q(\phi(z), y) = \varepsilon$. This means that the open ball $B(x, \delta)$ is contained in the set $\phi^{-1}(B(y, \varepsilon))$, so $\phi^{-1}(B(y, \varepsilon))$ is open.

To show the converse, let $\varepsilon > 0$. Then for every $x \in X$ the set $\phi^{-1}(B(\phi(x), \varepsilon))$ is open. As $x \in \phi^{-1}(B(\phi(x), \varepsilon))$, there exists a number $\delta > 0$ so that the ball $B(x, \delta)$ is contained in $\phi^{-1}(B(\phi(x), \varepsilon))$. This means that whenever z is in $B(x, \delta)$, then $\phi(z)$ is in $B(\phi(x), \varepsilon)$. This implies that ϕ is continuous, completing the proof. \square

On any metric space (X, d) one can define the Borel sets in X , denoted $\mathcal{B}(X)$, as the σ -algebra generated by the open sets of X . Given two metric spaces (X, d) and (Y, q) we define a **Borel measurable** transformation $T : X \rightarrow Y$ to be a transformation such that the T -inverse image of any Borel set in Y is a Borel set in X . Then we have the following lemma.

Lemma 3.12.4. *Let (X, d) and (Y, q) be two metric spaces and $T : X \rightarrow Y$ a transformation. If T is continuous, then it is Borel measurable.*

Proof. Let

$$\mathcal{A} = \{A \subset Y : \phi^{-1}(A) \in \mathcal{B}(X)\}.$$

Clearly \mathcal{A} contains the open sets of Y . Show that it is a monotone class and conclude that it contains $\mathcal{B}(Y)$ (Exercise 1). \square

Exercises

- (1) Complete the proof of Lemma 3.12.4.

- (2) Let Ω be the subset of Σ_3^+ consisting of all sequences x that do not have the word 010 at any place. Show that Ω is a closed subset of Σ_3^+ .
- (3) For $x, y \in \Sigma_N$ define $I(x, y) = \min\{|i| : x(i) \neq y(i)\}$. Define d on Σ_N by $d(x, y) = 2^{-I(x, y)}$, if $x \neq y$ and $d(x, x) = 0$. Show that d is a metric and Σ_N with this metric is compact for $N \geq 2$.

3.13. Symbolic Systems

We study two important transformations defined on symbolic spaces. The first is called the **shift**. Define $\sigma : \Sigma_N^+ \rightarrow \Sigma_N^+$ by shifting the entries of $x \in \Sigma_N^+$ to the left. More precisely, the i entry of $\sigma(x)$ is the $i + 1$ entry of x :

$$(\sigma(x))(i) = x(i + 1).$$

One can easily construct points whose positive orbit under σ is dense. In fact, let w be such that for each integer $k > 0$ it contains all words of length k . So, for example,

$$\begin{aligned} w &= 0100011011000001010011 \cdots, \\ \sigma^5(w) &= 11011000001010011 \cdots. \end{aligned}$$

It is clear that the positive orbit of w is dense, so σ is topologically transitive. However, it is not minimal as it has many periodic points. For example, $x = 10\overline{10} \cdots$ is a point of period 2.

The next map that we study is the **odometer map** $\tau : \Sigma_N^+ \rightarrow \Sigma_N^+$. This is defined by addition by 1 (mod N) to the first coordinate with a carry to the right. So, for example, if $x = 012\overline{012} \cdots$ in Σ_3^+ , then $\tau(x) = 112\overline{012} \cdots$, $\tau^2(x) = 212\overline{012} \cdots$, $\tau^3(x) = 022\overline{012} \cdots$.

We now define a probability measure on Σ_N^+ . We present the construction in the case when $N = 2$ as this contains all the ideas. The case for arbitrary N is left as an exercise. This measure will be defined on the Borel σ -algebra of Σ_N^+ .

First, we define a Borel measurable invertible transformation between a Borel subset of $[0, 1)$ and Σ_2^+ . As discussed in Section 2.2,

every number x in $[0, 1] \setminus D$ has a unique representation in binary form as

$$x = \sum_{i=1}^{\infty} \frac{x_i}{2^i},$$

where $x_i \in \{0, 1\}$ and D is the set of binary rational numbers. Let $I_0 = [0, 1] \setminus D$. For $x = \sum_{i=1}^{\infty} \frac{x_i}{2^i} \in I_0$ define the map $\psi : I_0 \rightarrow \Sigma_2^+$ by

$$(3.8) \quad \psi(x) = x_1 x_2 \cdots x_n \cdots .$$

Then ψ is a one-to-one and continuous map (Exercise 2). It follows that it is Borel measurable and can be used to define a measure ν on Σ_2^+ by

$$\nu(A) = \lambda(\psi^{-1}(A)),$$

for every A in $\mathcal{B}(\Sigma_2^+)$.

Lemma 3.13.1. *The set function ν is a probability measure defined on the Borel σ -algebra of Σ_2^+ . Furthermore, for any cylinder $[a_0 \cdots a_{k-1}]$,*

$$\nu([a_0 \cdots a_{k-1}]) = \frac{1}{2^k}.$$

Proof. If $\{A_n\}$ is a disjoint collection of Borel sets in Σ_2^+ , then $\{\psi^{-1}(A_n)\}$ are disjoint and Borel in I_0 and furthermore,

$$\begin{aligned} \nu\left(\bigsqcup_{n=1}^{\infty} A_n\right) &= \lambda \circ \psi^{-1}\left(\bigsqcup_{n=1}^{\infty} A_n\right) \\ &= \lambda\left(\bigsqcup_{n=1}^{\infty} \psi^{-1}(A_n)\right) \\ &= \sum_{n=1}^{\infty} \lambda(\psi^{-1}(A_n)) = \sum_{n=1}^{\infty} \nu(A_n). \end{aligned}$$

Also, $\nu(\Sigma_2^+) = \lambda(\psi^{-1}(\Sigma_2^+)) = \lambda(I_0) = 1$. □

Theorem 3.13.2. *Let (X, d) and (Y, q) be two metric spaces and $S : X \rightarrow Y$ a map. Let μ be a σ -finite Borel measure on $\mathcal{B}(X)$ and let ν be a σ -finite Borel measure on $\mathcal{B}(Y)$. If for every open set G in Y , $S^{-1}(G)$ is a Borel set in X and*

$$\mu(S^{-1}(G)) = \nu(G),$$

then S is Borel measurable and measure-preserving.

Proof. The proof is similar to the second proof of Theorem 3.4.1. Write $Y = \bigsqcup Y_n$, where the Y_n are Borel and of finite ν -measure. Set

$$\mathcal{A}_n = \{A : A \in \mathcal{S}(Y_n) \text{ and } T^{-1}(A) \in \mathcal{S}(X), \mu(T^{-1}(A)) = \nu(A)\}.$$

Showing that \mathcal{A}_n contains the Borel sets in Y_n , for each $n \geq 1$, is left to the reader. \square

We prove the following lemma for $N = 2$, with the general case left as an exercise.

Corollary 3.13.3. *The transformations σ and τ are both measure-preserving with respect to the Borel probability measure ν on Σ_2^+ .*

Proof. We first describe the value of the measure ν on cylinders. By definition, $\nu([0]) = \lambda(\psi^{-1}([0])) = \lambda([0, 1/2)) = 1/2$. In this way we see that

$$\begin{aligned} \nu([a_0 \cdots a_{k-1}]) &= \lambda(\psi^{-1}([a_0 \cdots a_{k-1}])) \\ &= \lambda\left(\left[\sum_{i=1}^k \frac{a_{i-1}}{2^i}, \sum_{i=1}^k \frac{a_{i-1}}{2^i} + \frac{1}{2^k}\right)\right) = \frac{1}{2^k}. \end{aligned}$$

Now,

$$\sigma^{-1}([a_0 \cdots a_{k-1}]) = [a_0 \cdots a_{k-1}0] \sqcup [a_0 \cdots a_{k-1}1].$$

So $\nu(\sigma^{-1}([a_0 \cdots a_{k-1}])) = \nu([a_0 \cdots a_{k-1}])$. As open sets in Σ_2^+ are countable unions of cylinders, Theorem 3.13.2 gives that σ is measure-preserving. The proof for τ is similar and left as an exercise. \square

Exercises

- (1) Show that the set of periodic points for $\sigma : \Sigma_N^+ \rightarrow \Sigma_N^+$ is dense.
- (2) Show that the map ψ in (3.8) is one-to-one and continuous.
- (3) Show that the odometer map τ is continuous and minimal.
- (4) Show that the shift map is continuous. (Hint: If x and y agree in k places, for some integer $k > 0$, then $\sigma(x)$ and $\sigma(y)$ agree in $k - 1$ places.)
- (5) Complete the proof of Corollary 3.13.3 by showing that τ is measure-preserving on cylinder sets.

- (6) Show that the odometer map τ on Σ_2^+ is isomorphic to the odometer map on $[0, 1)$.
- (7) Show that the shift map σ on Σ_2^+ is isomorphic to the doubling map.
- (8) Show that the odometer map is ergodic without appealing to Exercise 6.
- (9) Show that the shift map is ergodic without appealing to Exercise 7.
- (10) Formulate and prove a lemma analogous to Lemma 3.13.1 in the case of Σ_N .
- (11) A map $T : (X, d) \rightarrow (Y, q)$ of topological spaces is said to be a **homeomorphism** if it is invertible and both T and T^{-1} are continuous. Show that the shift map σ on Σ_N is a homeomorphism.

Open Question B. This is an open question due to Furstenberg. Let $T : [0, 1) \rightarrow [0, 1)$ be defined by $T(x) = 2x$ and let $S : [0, 1) \rightarrow [0, 1)$ be defined by $S(x) = 3x$. Observe that T and S are measure-preserving transformations with respect to Lebesgue measure. Also, the atomic measure supported at the point $\{0\}$ is also invariant for both T and S . There are uncountably many invariant nonatomic Borel measures that are invariant for T , and others that are invariant for S . The open question is whether there is any nonatomic Borel probability measure μ on $[0, 1)$, other than Lebesgue measure, such that μ is invariant for both T and S . This question remains open but special cases are known. a) Find several measures that are invariant for T alone and for S alone. b) Search the literature to find the special solutions to this question by Rudolph. c) Search the literature to find what is currently known about this question. d) Write a paper outlining the partial solution of Rudolph's and later developments of this solution. e) Write a paper describing the latest developments on this question. f) Solve the problem and tell the author.