Measure, Integrals, and Transformations: Lebesgue Integration and the Ergodic Theorem

Maxim O. Lavrentovich

November 15, 2007

Contents

0 Notation **1**

1 Introduction **2**

1.1 Sizes, Sets, and Things 2

1.2 The Extended Real Numbers 3

2 Measure Theory **5**

2.1 Preliminaries 5

2.2 Examples 8

2.3 Measurable Functions 12

3 Integration **15**

3.1 General Principles 15

3.2 Simple Functions 20

3.3 The Lebesgue Integral 28

3.4 Convergence Theorems 33

3.5 Convergence 40

4 Probability **41**

4.1 Kolmogorov's Probability 41

4.2 Random Variables 44

5 The Ergodic Theorem **47**

5.1 Transformations 47

5.2 Birkhoff's Ergodic Theorem 52

5.3 Conclusions 60

6 Bibliography **63**

0 Notation

We will be using the following notation throughout the discussion. This list is included here as a reference. Also, some of the concepts and symbols will be defined in subsequent sections. However, due to the number of different symbols we will use, we have compiled the more archaic ones here.

$\mathbb{R}, \mathbb{N}$:	the real and natural numbers
$\bar{\mathbb{R}}$:	the extended natural numbers, i.e. the interval $[-\infty, \infty]$
$\bar{\mathbb{R}}_+$:	the nonnegative (extended) real numbers, i.e. the interval $[0, \infty]$
Ω :	a sample space, a set of possible outcomes, or any arbitrary set
$\mathcal{F}$:	an arbitrary σ-algebra on subsets of a set Ω
$\sigma\langle\mathcal{A}\rangle$:	the σ-algebra generated by a subset $\mathcal{A} \subseteq \Omega$.
T :	a function between the set Ω and Ω itself
ϕ, ψ :	simple functions on the set Ω
μ :	a measure on a space Ω with an associated σ-algebra $\mathcal{F}$
(X_n) :	a sequence of objects X_i, i. e. $X_1, X_2, \ldots$
$f_n \to f$:	a sequence of functions (f_n) converging pointwise to f
$A_n \uparrow A$:	here A is a set and (A_n) is a sequence of sets and $\bigcup_{i=1}^{\infty} A_i = A$ and $A_1 \subseteq A_2 \subseteq \ldots$
$A_n \downarrow A$:	as above, replacing $\subseteq$ with $\supseteq$ and $\bigcup_{i=1}^{\infty}$ with $\bigcap_{i=1}^{\infty}$
$f_n \uparrow f$:	a sequence of functions $(f_n : D \to \mathbb{R})$ and a function $f : D \to \mathbb{R}$ such that $f_1(x) \leq f_2(x) \leq \ldots$ for all $x \in D$ and $f_n \to f$
$f_n \downarrow f$:	as above, replacing $\leq$ with $\geq$
$A^{\mathcal{C}}$:	the complement of A with respect to Ω, i.e. the set of all $x \in \Omega$ such that $x \notin A$
$A \setminus B$:	the set of all points in A but not in B

1 Introduction

1.1 Sizes, Sets, and Things

The primary objective of this paper is to study the structure of spaces that have "sizes" associated with them. For example, we will be thinking about lengths, areas, and volumes. We want to formulate a mathematical formalism that will allow us to define the "size" of a set in a very abstract way. Now, we know intuitively that a size will be some real number, and that there are many ways in which we can assign a size to an object. Thus, it is clear that the two primary fields of mathematics that we will apply to build our formalism are the study of the real numbers, or real analysis, and the study of sets, or set theory. Indeed, in order to talk about sizes abstractly, we have to be able to look at objects and regions as abstract entities. Set theory provides us with this language and we will assume that the basic tenets of set theory are familiar to the reader. However, we will define a few key concepts that we will be using throughout the discussion.

Definition 1.1.1. Let $\{A_i \mid i \in \Lambda\}$ be an arbitrary collection of sets. We say that these sets are *mutually* or *pairwise disjoint* if for any $m, n \in \Lambda$ such that $n \neq m$, $A_n \cap A_m = \emptyset$.

Definition 1.1.2. We say that a sequence of sets $(A_i \mid i \in \mathbb{N})$ is *decreasing* if $A_1 \supseteq A_2 \supseteq \ldots$. Now, if we have a set B such that $B = \bigcap_{n=1}^{\infty} A_n$ for a decreasing sequence (A_i), then we write $A_n \downarrow B$. Similarly, a sequence of sets (A_n) is *increasing* if $A_1 \subseteq A_2 \subseteq \ldots$. Moreover, if $B = \bigcup_{n=1}^{\infty} A_n$ for such a sequence, then we write $A_n \uparrow B$.

Now, we also know that sizes are real numbers. If we think about the area of a region on a plane, say, we think about some value of the area in square centimeters, for instance. Also, if we are talking about the number of elements in a finite set, this number is some integer, which is included in the real numbers. Finally, if we want to talk about the probability of a certain event, we think of a real number between zero and one. The point is, all of these examples are examples of sizes or measures of sets. So, we see that a measure will have to yield a real number. Therefore, it will be important for us to have a solid background in real analysis while constructing our formalism. Again, we will assume that the reader is familiar with the basic ideas in this field. However, some of the notations and definitions we have given for sets above, have corresponding concepts for sequences of real numbers and real-valued functions. Specifically, we can say that for a sequence of real numbers (a_n), the notation $a_n \uparrow a$ for some $a \in \mathbb{R}$ means that a_n converges to a and $a_1 \leq a_2 \leq a_3 \leq \ldots$. We

have an analogous definition for $a_n \downarrow a$. Thus, a similar idea can be applied to real valued functions. Consider a sequence of functions (f_n), where each function maps elements in some set Ω to the real numbers $\mathbb{R}$. Then, $f_n \uparrow f$ means that $f_1(x) \leq f_2(x) \ldots$ for all $x \in \Omega$, and f_n converges to f pointwise. Note that we are assuming here that the reader is familiar with the notions of pointwise and uniform convergence of functions from elementary real analysis. We will define some more concepts concerning these issues in the next subsection.

Finally, in the proofs that follow it will be handy to recall the DeMorgan's laws for sets. We will use the language employed in the introductory real analysis book by Schumacher [Sch].

Theorem 1.1.3. (*DeMorgan's laws*) *Let Ω be a set. Let $\{B_\alpha\}_{\alpha\in\Lambda}$ be a collection of subsets of Ω. Then, we have that*

$$1. \quad \left(\bigcup_{\alpha\in\Lambda} B_\alpha\right)^{\mathcal{C}} = \bigcap_{\alpha\in\Lambda} B_\alpha^{\mathcal{C}}$$

$$2. \quad \left(\bigcap_{\alpha\in\Lambda} B_\alpha\right)^{\mathcal{C}} = \bigcup_{\alpha\in\Lambda} B_\alpha^{\mathcal{C}}$$

1.2 The Extended Real Numbers

In order to facilitate the subsequent discussions, we want to be able to talk about not only the real line, but also the points $+\infty$ and $-\infty$. These two points are not points in the real numbers, but rather they represent points that satisfy the following relation:

$$-\infty < x < \infty \text{ for all } x \in \mathbb{R}.$$

We can see that the relation above makes intuitive sense because indeed, we cannot have a real number that is less than negative infinity and greater than positive infinity. We will also *define* operations between the infinities and the real numbers in the following way, following [Ad] pg 53:

Definition 1.2.1. Let $x \in \mathbb{R}$. We define operations on $\mathbb{R} \cup \{-\infty, \infty\}$ in the following way:

i. $x + (\pm\infty) = \pm\infty + x = \pm\infty$

ii. $x(\pm\infty) = \pm\infty$ if $x > 0$

iii. $x(\pm\infty) = \mp\infty$ if $x < 0$

The other possibilities, such as division by the infinities, we will not define. We will also not consider adding a positive infinity to a negative infinity. However, the above definitions should be intuitively reasonable. We are simply saying that multiplication or addition of a finite real number x by an infinite number yields an infinite number. The set $\mathbb{R} \cup \{-\infty, +\infty\}$ is referred to as the *extended* real numbers. In using the extended real numbers, we keep in mind that the two infinities are not real numbers. They are simply useful conventions that allow us to talk about infinite sizes, such as the size of the interval $[1, \infty)$. From now on, whenever we mention real numbers, we will be referring to the extended real numbers. Also, we will denote our extended real numbers by the symbol $\bar{\mathbb{R}}$.

In this paper we will also be concerned with not only sequences of sets, but also sequences of (extended) real numbers and real valued functions. Therefore, we have similar definitions for the limits of such sequences. We recall from basic analysis, the following terminology([Sch] pg. 330):

Definition 1.2.2. Let (a_n) be a bounded sequence of real numbers. We define two sequences based on the sequence (a_n). The *upper sequence* is defined as $\bar{a}_n = \sup_{k \geq n} a_k$. Similarly, the *lower sequence* is $\underline{a}_n = \inf_{k \geq n} a_k$. From the definitions of the supremum and infimum, it is easy to show that the upper sequence and the lower sequence are monotonic. Therefore, the limits of these two sequences must exist, and we call these limits the supremum and the infimum limits, respectively, of the sequence (a_n). These are denoted by $\limsup a_n$ and $\liminf a_n$.

Now, suppose that we have a sequence of real numbers (a_n) that is *not* bounded. A fundamental result in analysis is that the limits $\limsup a_n$ and $\liminf a_n$ exist, provided that we allow them to be $\pm\infty$. Also, we will often use the notation $\limsup_{n\to\infty} a_n$ when we do not want to be ambiguous about the index of the sequence for which we are evaluating the supremum and infimum limits. All of these definitions can be generalized to sequences of real-valued *functions* on some set Ω. We simply look at the real number sequences that are created when we evaluate the functions at particular points on their domains. In other words, for a particular sequence of functions (f_n) (where $f_n : \Omega \to \mathbb{R}$ for each $n \in \mathbb{N}$), for each $x \in \Omega$, the sequence $(f_n(x))$ is a sequence of real numbers. Therefore, we can define the functions $\limsup f_n$ and $\liminf f_n$ at each point $x \in \Omega$ by using the definitions we have above for sequences of real numbers. We may also conclude that such limiting functions ($\limsup f_n$ and $\liminf f_n$) exist for all real-valued functions. We also recall from analysis that a sequence of real-valued functions (f_n) converges pointwise to some function f if and only if $\limsup f_n = \liminf f_n$.

2 Measure Theory

2.1 Preliminaries

As we mentioned previously, measure theory is the study of the sizes of sets. Specifically, we usually look at some large set Ω, and then we compute the sizes of subsets of Ω. For example, a particular Ω might be all of space $\mathbb{R}^3$. Then, we can think about volumes of particular regions (subsets) in $\mathbb{R}^3$. However, the subsets of Ω that we can actually measure must have certain natural properties. These natural properties motivate the definition of σ-algebra, which is basically a collection of subsets of Ω that behave in a nice way. Thus, using the language found on page 9 in Athreya's book [Ath],

Definition 2.1.1. A collection of sets $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called a *σ-algebra* if

i. $\Omega \in \mathcal{F}$

ii. $A \in \mathcal{F}$ implies $A^{\mathcal{C}} \in \mathcal{F}$

iii. $A, B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$

iv. $A_n \in \mathcal{F}$ for $n \geq 1$ implies $\bigcup_{n \geq 1} A_n \in \mathcal{F}$

In other words, a σ-algebra is a class of subsets of Ω that contains Ω and is closed under complementation and countable unions.

Given this definition, we will now use the symbol $\mathcal{F}$ to refer to the σ-algebra associated with the set Ω. Let us now consider the different conditions for a σ-algebra. The first condition basically says that the whole set we are looking at has to be in the σ-algebra. This makes sense because we define Ω in such a way that the structure of the set itself allows it to be measured. We know, for example, that there are a variety of metrics associated with the three-dimensional real space R^3. The second condition means that for every nicely behaved set $A \in \mathcal{F}$, we can think about the set of elements not in the set. This set should also be nicely behaved because we want to be able to consider such things like the probability of an event *not* occurring, or the volume outside of some region in real space. The last two events simply say that we can add together nicely behaved sets to get another nicely behaved set. For example, if we have two regions in space whose volume can be measured, then certainly we expect that the entire region the two regions cover can be measured, as well. Now

that we have tried to clear up these issues a little bit, we will prove an easy corollary of the definition of σ-algebra given above.

Corollary 2.1.2. *Suppose (A_n) is a sequence of subsets of Ω such that $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$. Then, the countable intersection $\bigcap_{n=1}^{\infty} A_n$ is also in $\mathcal{F}$.*

Proof. Since $\mathcal{F}$ is closed under complementation, $A_n^{\mathcal{C}} \in \mathcal{F}$ for all $n \in \mathbb{N}$. Moreover, because $\mathcal{F}$ is closed under countable unions, $\bigcup_{n=1}^{\infty} A_n^{\mathcal{C}} \in \mathcal{F}$. Finally, by DeMorgan's Laws,

$$\bigcup_{n=1}^{\infty} A_n^{\mathcal{C}} = \left(\bigcap_{n=1}^{\infty} A_n \right)^{\mathcal{C}} \in \mathcal{F} \;\Rightarrow\; \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}.$$

□

Perhaps the most trivial example of a σ-algebra corresponding to a set Ω is just the collection $\{\emptyset, \Omega\}$. Notice that this collection satisfies all the properties given above. Indeed, $\emptyset \cap \Omega = \emptyset$ and $\emptyset \cup \Omega = \Omega$. Thus, this is a very easy example. Also, notice that this example is necessarily the *smallest* σ-algebra possible for a set. Another easy example of a σ-algebra is just the powerset $\mathcal{P}(\Omega)$ of Ω. Recall that a powerset is simply the collection of all subsets of Ω. Thus, it must satisfy our conditions because any countable unions, complements, and things like that are clearly subsets of Ω, and thus must be included in the powerset. Also, since the powerset includes all possible subsets of Ω, it is necessarily the *largest* possible σ-algebra. It is also possible, given some collection $\mathcal{A}$ of subsets of Ω, to construct the smallest σ-algebra containing $\mathcal{A}$. We will now describe how to do this by employing the language found on page 11 in Athreya's book [Ath].

Definition 2.1.3. If $\mathcal{A}$ is a class of subsets of Ω, then the σ-algebra generated by $\mathcal{A}$, which we will denote by $\sigma\langle\mathcal{A}\rangle$, is

$$\sigma\langle\mathcal{A}\rangle \equiv \bigcap_{\mathcal{F} \in \mathcal{I}(\mathcal{A})} \mathcal{F},$$

where $\mathcal{I}(\mathcal{A}) \equiv \{\mathcal{F} \mid \mathcal{A} \subset \mathcal{F} \text{ and } \mathcal{F} \text{ is a } \sigma\text{-algebra on } \Omega\}$ is the collection of all the σ-algebras containing our collection $\mathcal{A}$. Notice that this definition is well defined because we know that the powerset $\mathcal{P}(\Omega)$ contains our collection $\mathcal{A}$ and is a σ-algebra. Thus, our set $\mathcal{I}(\mathcal{A})$ is not empty and we can take the intersection. Finally, notice that $\sigma\langle\mathcal{A}\rangle$ is a σ-algebra because given any sequence of sets $A_1, A_2, \ldots, A_n \in \sigma\langle\mathcal{A}\rangle$, then their intersections, complements, and unions must also be in $\sigma\langle\mathcal{A}\rangle$ because from the properties of σ-algebras, we know that these intersections, complements, and unions will be in every single σ-algebra containing $\mathcal{A}$.

The most important example of a generated σ-algebra is the one that is generated by the open sets in $\mathbb{R}^n$. So, we have the following definition.

Definition 2.1.4. The *Borel σ-algebra* on $\mathbb{R}^n$ is defined as the σ-algebra generated by the collection of open sets in $\mathbb{R}^n$.

We now want to be able to characterize the size of particular subsets of our set Ω. Indeed, we want to be able to assign numbers to elements in $\mathcal{F}$. Therefore, we want to consider functions that map elements in $\mathcal{F}$ to the real numbers. In order for these functions to reasonably represent the "size" of the set, we need to define certain natural properties for these functions. Further, we shall now call these functions, that assign sizes to sets according to the properties that follow, *measures*. So, from Athreya's book on page 14,

Definition 2.1.5. Consider a (extended) real-valued function μ on a σ-algebra $\mathcal{F}$ on a set Ω. We call μ a measure if

i. $\mu(F) \in [0, \infty]$ for all $A \in \mathcal{F}$

ii. $\mu(\emptyset) = 0$

iii. for any pairwise disjoint collection of sets $A_1, A_2, \ldots \in \mathcal{F}$ with $\bigcup_{n\geq 1} A_n \in \mathcal{F}$,

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n).$$

We call the triplet $(\Omega, \mathcal{F}, \mu)$ a *measure space*.

First, notice that the conditions enumerated above work well with the conditions we had for our σ-algebra. Indeed, the definitions of measure and σ-algebra allow us to add and subtract subsets of Ω and have the sizes of these subsets behave nicely as we manipulate them. Anyway, let us consider Def. 2.1.5 by examining each condition. The first condition says that this function must assign a number from 0 to ∞ to each set in $\mathcal{F}$. This means that we cannot assign a negative number to a set. Intuitively, this makes sense because a size is intrinsically a positive number. We cannot think about the area of some region in the plane, the probability of some event, or the mass of gold in a piece of ore as a negative number. The next condition simply states that if we are measuring a set

with no elements, i.e. an empty set, then we should get a size of zero. Finally, the third condition says that if we have multiple, separate, distinct regions we want to measure, such as different pieces of ore, or different areas in the plane, or different outcomes for a random event, then the combined size of the regions is simply the sum of the sizes of the individual regions. Thus, we see that these three conditions provide a mathematical foundation for our intuition about "size".

2.2 Examples

We will now present concrete illustrations of the general concepts described in the previous sections. The purpose of the first example is to illustrate the definition of a measure at a very basic level. So, suppose we have a finite set $\Omega = \{a, b\}$ with just two elements. Here, a and b can be whatever we like, like an "a"pple and a "b"anana, for example. We may explicitly construct the power set of Ω. This will be our σ-algebra. We write that $\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{a\}, \{b\}, \Omega\}$. Now, consider a function $\mu : \mathcal{P}(\Omega) \to \mathbb{R}$ constructed in the following way: $\mu(\emptyset) = 0$, $\mu(\{a\}) = \mu(\{b\}) = 1$, and $\mu(\Omega) = 2$. From this construction, we see that the function μ is simply counting the number of elements in subsets of Ω. This corresponds well with our intuition about the "size" of a finite set of arbitrary elements. Not surprisingly, then, we find that μ satisfies the properties of a measure given in Def. 2.1.5. Specifically, notice that the first and second properties follow trivially from the construction of μ. We can show that the third property is satisfied by writing down all of the pairwise disjoint collections of subsets (with more than one element) of Ω: $\emptyset$ and any other subset; $\emptyset$, $\{a\}$, $\{b\}$; and $\{a\}$, $\{b\}$. Then, $\mu(\emptyset \cup \{a\}) = \mu(\{a\}) = 1 = \mu(\{a\}) + \mu(\emptyset)$, with a similar result for $\emptyset \cup \{b\}$ and $\emptyset \cup \Omega$. Also, $\mu(\emptyset \cup \{a\} \cup \{b\}) = \mu(\{a\} \cup \{b\}) = \mu(\Omega) = 2 = \mu(\{a\}) + \mu(\{b\})$, and so on. Consequently, we see that this very simple function satisfies our conditions for a measure in Def. 2.1.5. This is why such a function is called a *counting measure*. Further, this function can be generalized to any finite or countably infinite set Ω.

Perhaps the most famous example of a measure is the "Lebesgue measure" on the real numbers. In order to get at this measure, however, we have to first look at the "Lebesgue outer measure". The latter concept was introduced by Lebesgue in 1902 and is based on covering subsets of the real numbers with a countably infinite number of intervals (pg. 23 [Bu]). This outer measure is defined on page 33 in Wheeden's book [Wh] and on page 131 in Asplund's textbook on integration [As]. We will roughly follow the definitions in these two sources.

Definition 2.2.1. Consider the real numbers $\mathbb{R}$ and open intervals $I = (a, b)$, where $a \leq b$ are two

real numbers. We say that the length of such an interval is $|b-a|$. Now, let us suppose we have some arbitrary subset of the real numbers $A \subseteq \mathbb{R}$. To define our Lebesgue outer measure, let us cover A with a countably infinite collection S of open intervals I_k, i.e. S is an open cover of A. Explicitly, $S = \{I_i\}_{1 \leq i \leq \infty}$, where $I_i = (a_i, b_i)$. Then, we can think about the sum

$$\sigma(S) = \sum_{i=1}^{\infty} |b_i - a_i|.$$

From the properties of the absolute value, it is clear that $0 \leq \sigma(S) \leq \infty$. Finally, we call the Lebesgue outer measure of A, which we will denote by $\mu_L^*(A)$, the quantity

$$\mu_L^*(A) = \inf \sigma(S),$$

where the infimum is taken over all open covers S of the set A. We see that $0 \leq \mu_L^*(A) \leq \infty$. Moreover, since the empty set is a subset of any set of real numbers, our infimum $\sigma(S)$ must include intervals with zero length. Therefore, the Lebesgue outer measure of the empty set must be zero.

The definition of the outer measure (Def. 2.2.1) can be phrased in terms of any kind of interval: closed, open, or neither. This is true because the outer measure is an infimum of sums of lengths of open intervals. Hence, we can just as easily cover a subset $A \subset \mathbb{R}$ with closed (or neither open nor closed) intervals and have the outer measure of A be the same because our closed intervals can be subsets of open intervals that also cover A and have lengths arbitrarily close to the lengths of the closed intervals. Moreover, Def. 2.2.1 can easily be generalized to $\mathbb{R}^n$. There, we consider n-dimensional intervals $I = \{\mathbf{x} : a_j < x_j < b_j, j = 1, 2, \ldots, n\}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is a vector. Then, instead of computing the *length* of this interval, we calculate the *volume*: $v(I) = \prod_{j=1}^{n} |b_j - a_j|$.

However, there is a reason we do not call the Lebesgue outer measure just a measure. The problem with this characterization of subsets of $\mathbb{R}$ is that it does not necessarily satisfy the third property of a measure given in Def. 2.1.5. The closest we can come to this third condition is the following property, which is called the *countable subadditivity* condition ([Wh] pg. 34):

Theorem 2.2.2. *If $E = \bigcup_{i=1}^{\infty} E_i$ is a countably infinite union of subsets of $\mathbb{R}$, then $\mu_L^*(E) \leq \sum_{i=1}^{\infty} \mu_L^*(E_i)$.*

Proof. We have two cases. First, if $\mu_L^*(E_i) = \infty$ for any $i \in \mathbb{N}$, then the theorem follows trivially from the nonnegativity of the Lebesgue outer measure. Thus, suppose that $\mu_L^*(E_i) < \infty$ for all $i \in \mathbb{N}$.

Let $\epsilon > 0$ be some arbitrarily tolerance. Given an index $k \in \mathbb{N}$, choose intervals $I_{k,j} = (a_{j,k}, b_{j,k})$ such that $E_k \subset \bigcup_{j=1}^{\infty} I_{k,j}$ and $\mu_L^*(E_k) \leq \sum_{j=1}^{\infty} |b_{j,k} - a_{j,k}| < \mu_L^*(E_k) + \epsilon\, 2^{-k}$. This is possible because $\mu_L^*(E_k)$ is an infimum of "lengths" of open covers (see Def. 2.2.1). Next, since $E \subset \bigcup_{j=1}^{\infty} \bigcup_{k=1}^{\infty} I_{k,j}$, we also know that $\mu_L^*(E) \leq \mu_L^* \left(\bigcup_{j=1}^{\infty} \bigcup_{k=1}^{\infty} I_{k,j} \right)$, because any open cover of the double union of the intervals $I_{i,k}$ must cover E, as well. Finally, we combine these results to find that

$$\mu_L^*(E) \leq \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |b_{j,k} - a_{j,k}| \leq \sum_{k=1}^{\infty} \left(\mu_L^*(E_k) + \frac{\epsilon}{2^k} \right) = \sum_{k=1}^{\infty} \mu_L^*(E_k) + \epsilon,$$

where we have exchanged the order of the two summations and summed the geometric series. The theorem follows because our ϵ was arbitrary. □

Unfortunately, the converse of Thm. 2.2.2 is not true in general. However, we can define a class of subsets of the real numbers whose Lebesgue outer measure satisfies all of the requirements of the measure given in Def. 2.1.5. These subsets are called *measurable*. We will now present a definition of this class of subsets, using the language in Wheeden's book on page 37 [Wh].

Definition 2.2.3. A subset A of $\mathbb{R}$ (or $\mathbb{R}^n$) is said to be *Lebesgue measurable*, or simply *measurable*, if given any $\epsilon > 0$, there exists an open subset B of the real numbers such that

$$A \subset B \text{ and } \mu_L^*(B \setminus A) < \epsilon.$$

If A is measurable, its Lebesgue outer measure is called its *Lebesgue measure*, and is denoted by $\mu_L(A)$.

Using this definition, it is now possible to show a variety of nice properties of measurable sets. However, the proofs of these properties are not particularly instructive; they simply rehash ideas from real analysis. Thus, we refer the reader to the presentation found on pages 37 through 42 in Wheeden's book [Wh]. We will just list the results.

Consider subsets of the real numbers. We can derive the following properties using Def. 2.2.3 and basic considerations from real analysis:

- Every open or closed subset is measurable.
- Every subset of outer measure zero is measurable.

- Every interval is measurable.
- Any countable intersection or union of measurable subsets is measurable.
- The Lebesgue measure satisfies the properties of a measure.
- The collection of measurable subsets of $\mathbb{R}$ (or $\mathbb{R}^n$) is a σ-algebra!
- Every set in the Borel σ-algebra is measurable (see Def. 2.1.4).

The last few results are especially important because they allow us to extend the general formalism we will develop for measures on σ-algebras in later sections to the Lebesgue measure on σ-algebras of measurable subsets of the real numbers.

Before we move on to other considerations, it is instructive to think about sets of measure zero. Consider that a set with a single real number, such as $\{a\}$, where $a \in \mathbb{R}$, has a Lebesgue outer measure of zero ($\mu_L^*(\{a\}) = |a - a| = 0$). This means, from the results given above, that every set with a single real number has a Lebesgue measure of zero. Now, suppose we have a pairwise disjoint collection of these single element sets. That is, we have a countably infinite number of single element sets A_i, where all of the elements are distinct real numbers: $A_i = \{a_i : a_i \in \mathbb{R}\}$ for all $i \in \mathbb{N}$ such that $a_i \neq a_j$ whenever $i \neq j$. From the third property of the measure in Def. 2.1.5, we conclude that $\mu_L(\bigcup_i A_i) = \sum_i \mu_L(A_i) = 0$. Hence, every countable subset of the real numbers has measure zero! This makes intuitive sense because we know that no matter how many distinct points we remove from the real line, we will still have uncountably many remaining numbers. Thus, we expect that the "length", or "size", of a countable number of points on the real line is zero. Interestingly, not all subsets of the real numbers with measure zero are countable! Perhaps the most famous example of such a subset was developed by Georg Cantor at the turn of the last century (pg. 252 [Bu]). A schematic of the construction of this subset is shown in Fig. 1. We will not prove that this set has measure zero, but refer the reader to Appendix A in Burk's book [Bu] for an exploration of the fascinating properties of this set.

The purpose of the previous discussions is to illustrate that the notion of measurability is not trivial. Indeed, we may not be able to measure certain sets at all! For example, in 1905, Vitali discovered a set of real numbers that is not Lebesgue measurable (pg. 266 [Bu]). This is why we went through the trouble discussing the distinction between the Lebesgue outer measure and measure on the real numbers. Also, in the discussions that follow, we will often prove results that are true

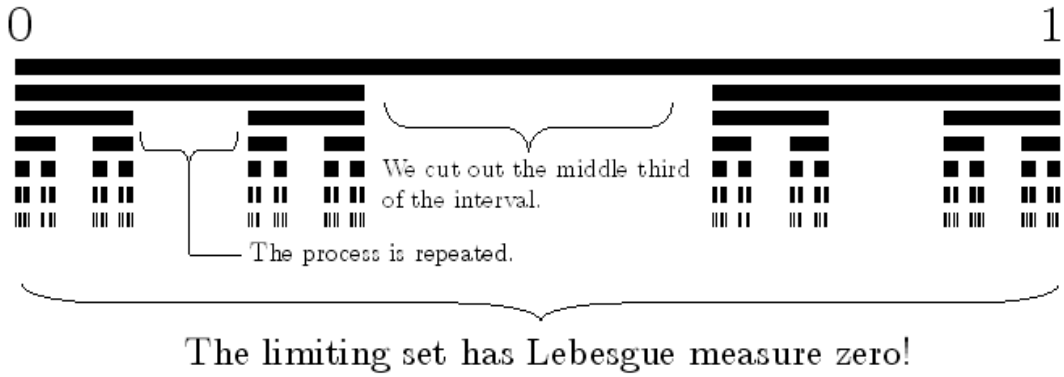

Figure 1: *This figure illustrates a subset of the* $[0, 1]$ *interval of the real numbers that is not countable, but has Lebesgue measure zero! It is called the Cantor middle third set. Part of this figure was provided by [Wi].*

for an entire set Ω, with the possible exception of some subset of Ω with measure zero. We will say that such results hold "almost everywhere", or for "almost every" point in the set. So, although we may speak casually about such a condition in the future, we realize that these subsets of measure zero may be highly complex and interesting (like the Cantor set). Their exclusion in a proof reflects the nontrivial fact that some results in measure and integration theory do not apply to pathological cases.

2.3 Measurable Functions

Sometimes, when we do not have a measure on a set Ω, we want to still be able to characterize certain functions defined on Ω. In this spirit, we will consider just sets Ω with corresponding σ-algebras $\mathcal{F}$. We will also keep in mind that whenever we refer to a function, we mean a real-valued function on the set Ω. Thus, we have the following definitions.

Definition 2.3.1. Let Ω be a nonempty set and let $\mathcal{F}$ be a σ-algebra on Ω. Then the pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

We may now define a special class of real-valued functions on the set Ω ($f : \Omega \to \mathbb{R}$).

Definition 2.3.2. The function $f : \Omega \to \mathbb{R}$ is *measurable* if for all $x \in \mathbb{R}$, the set $f^{-1}((-\infty, x)) \equiv \{y \in \Omega \mid f(y) < x\}$ is an element of $\mathcal{F}$.

Now, what the previous definition is saying is that f behaves well on its domain. That is, if we look at any $x \in \mathbb{R}$, and consider the subset $(-\infty, x)$ of the range of f, then we know that the inverse image of this subset of the range is an element in $\mathcal{F}$, i.e. a reasonable region to which we can assign a size. These measurable functions have some nice properties. We will prove them in the theorems that follow.

The first objective is to show that the inequality in Def. 2.3.2 is arbitrary. By allowing us to define a measurable function in many different ways, we will be able to derive some nice behaviors of these functions. The proof of the following lemma follows the one outlined by Adams on page 54 of his book [Ad].

Lemma 2.3.3. *The following are equivalent for all $x \in \mathbb{R}$:*

i. $\{y \in \Omega : f(y) \leq x\} \in \mathcal{F}$

ii. $\{y \in \Omega : f(y) > x\} \in \mathcal{F}$

iii. $\{y \in \Omega : f(y) \geq x\} \in \mathcal{F}$

iv. $\{y \in \Omega : f(y) < x\} \in \mathcal{F}$

Proof.
$(i.) \Leftrightarrow (ii.)$: We know that the set $\{y \in \Omega : f(y) > x\}$ is the complement of the set $\{y \in \Omega : f(y) \leq x\}$. Thus, since $\mathcal{F}$ is a σ-algebra and closed under complementation (Thm. 2.1.1), we conclude that if $\{y \in \Omega : f(y) > x\} \in \mathcal{F}$, then $\{y \in \Omega : f(y) \leq x\} \in \mathcal{F}$, as well. The converse is true by the same argument.

$(iii.) \Leftrightarrow (iv.)$: Once again, the two sets in question are complementary. The implications follow by the same argument given above.

$(ii.) \Rightarrow (iii.)$: We will first show that

$$\{y \in \Omega : f(y) \geq x\} = \bigcap_{i=1}^{\infty} \left\{ y \in \Omega : f(y) > x - \frac{1}{n} \right\}.$$

$(\subseteq)$: Let $p \in \{y \in \Omega : f(y) \geq x\}$. In other words, $f(p) \geq x$. Then, we know that this means that for all $n \in \mathbb{N}$, $f(p) \geq x > x - \frac{1}{n}$. Therefore, $p \in \bigcap_{i=1}^{\infty} \left\{ y \in \Omega : f(y) > x - \frac{1}{n} \right\}$.

($\supseteq$): Let $p \in \bigcap_{n=1}^{\infty} \left\{y \in \Omega \,:\, f(y) > x - \frac{1}{n}\right\}$. Now, consider an arbitrary $\epsilon > 0$. Since $f(p) > x - 1/n$ for all $n \in \mathbb{N}$, choose $M \in \mathbb{N}$ such that $1/M < \epsilon$. We conclude that $f(p)+\epsilon > f(p)+1/M > x$. From the basic properties of the real numbers, this means that $f(p) \geq x$ and so $p \in \{y \in \Omega \,:\, f(y) \geq x\}$.

Considering the equality we proved above, we realize that the sets $\left\{y \in \Omega \,:\, f(y) > x - \frac{1}{n}\right\}$ must be in $\mathcal{F}$ for all $n \in \mathbb{N}$ by our hypothesis (ii). Therefore, since $\mathcal{F}$ is closed under countable intersections (Corr. 2.1.2), $\bigcap_{n=1}^{\infty} \left\{y \in \Omega \,:\, f(y) > x - \frac{1}{n}\right\} = \{y \in \Omega \,:\, f(y) \geq x\} \in \mathcal{F}$, as well.

$(iv.) \Rightarrow (i.)$: This implication must follow by a parallel proof to the one given above for $(ii.) \Rightarrow (iii.)$. □

Now that we have established the many ways in which a function can be measurable, we can derive a few nice properties of measurable functions. The first property, which will be stated without proof, is that any finite sum of measurable functions is a measurable function. Also, any finite product of measurable functions is a measurable function. These two properties follow from the definitions given above and are proven on page 42 in [Ath]. We will omit the proofs in favor of showing explicitly the following properties of sequences of measurable functions (outlined on page 135 in Burk [Bu]).

Theorem 2.3.4. *Suppose (f_k) is a sequence of measurable functions on Ω. Then, the following functions are also measurable:*

i. $\bar{f}_k = \sup\{f_k, f_{k+1}, f_{k+2}, \ldots\}$ *and* $\underline{f}_k = \inf\{f_k, f_{k+1}, f_{k+2}, \ldots\}$ *for* $k = 1, 2, \ldots$

ii. $\limsup f_k = \lim \bar{f}_k$ *and* $\liminf f_k = \lim \underline{f}_k$

iii. If $\lim f_k$ *converges pointwise to a function f everywhere in Ω, then f is also measurable*

Proof.
$(i.)$: First we will show that $\{y \in \Omega \mid \bar{f}_k(y) > x\} = \bigcup_{n=k}^{\infty} \{y \in \Omega \mid f_k(y) > x\}$.
($\subseteq$): We will proceed by contrapositive. Pick some arbitrary $p \in \Omega$. Now, suppose that $p \notin \bigcup_{n=k}^{\infty} \{y \in \Omega \mid f_k(y) > x\}$. Consequently, $f_n(p) \leq x$ for all $n \geq k$. However, this means that x is an upper bound for the set $\{f_k(p), f_{k+1}(p), \ldots\}$. By the definition of a supremum, $\bar{f}_k(p) \leq x$. We find that $p \notin \{y \in \Omega \mid \bar{f}_k(y) > x\}$. We are now done.
($\supseteq$): Suppose that $p \in \bigcup_{n=k}^{\infty} \{y \in \Omega \mid f_k(y) > x\}$. Thus, there exists some $l \geq k$ such that $f_l(p) > x$. However, by the definition of a supremum, we now see that $\bar{f}_k(p) \geq f_l(p) > x$, as well. We conclude that $p \in \{y \in \Omega \mid \bar{f}_k(y) > x\}$.

The equality has now been shown and we reason that $\{y \in \Omega \mid \bar{f}_k(y) > x\} = \bigcup_{n=k}^{\infty}\{y \in \Omega \mid f_k(y) > x\} \in \mathcal{F}$ since $\mathcal{F}$ is closed under countable unions and each function f_k is measurable. Thus, by Thm. 2.3.3, $\bar{f}_k$ must be a measurable function. For $\underline{f}_k(y)$, we realize that $\{y \in \Omega \mid \underline{f}_k(y) < x\} = \bigcup_{n=k}^{\infty}\{y \in \Omega \mid f_k(y) < x\} \in \mathcal{F}$ by a parallel proof to the one give above for $\bar{f}_k(y)$. By Thm. 2.3.3, we now know that $\underline{f}_k(y)$ is measurable.

(*ii.*): Consider that part (*i.*) tells us that $(\bar{f}_k)$ and $(\underline{f}_k)$ are, respectively, nonincreasing and nondecreasing sequences of measurable functions. Consequently, by the respective nonincreasing and nondecreasing properties of these sequences, we write that $\limsup f_k = \lim_{k\to\infty} \bar{f}_k = \inf\{\bar{f}_1, \bar{f}_2, \ldots\}$ and $\liminf f_k = \lim_{k\to\infty} \underline{f}_k = \sup\{\underline{f}_1, \underline{f}_2, \ldots\}$. We again use part (*i.*) to conclude that $\limsup f_k$ and $\liminf f_k$ must be measurable.

(*iii.*): Suppose that $\lim f_k$ converges pointwise to a function f everywhere in Ω. From our considerations about real-valued functions, we have that $\limsup f_k = \liminf f_k = f$. From the parts above, f must be measurable. □

Let us consider this result for a moment. It is very important. We already know from elementary real analysis that pointwise convergence of functions is, in a sense, the "weakest" kind of convergence. For example, we know that if a sequence of functions (f_n) converges uniformly to a function f, then the sequence must converge pointwise to f, as well. Thus, this result is powerful because we can deduce the measurability of a function by simply constructing a sequence of functions that eventually approach the function of interest at any given point on its domain. Finally, we keep in mind that the objective here is to build up a formal mathematical structure that will allow us to characterize subsets of sets Ω. Indeed, we want to find a class of nicely behaved, real-valued functions on these subsets. As we shall see, the purpose here is to find functions that are amenable to integration.

3 Integration

3.1 General Principles

We now want to build up the mathematical formalism that will allow us to integrate measurable functions on Ω. Recall that Ω can be a variety of sets, including the real numbers, or a space of possible outcomes, or even the set of all the words in this paper! Regardless, we see that the notion

of this set Ω is abstract. Thus, it is difficult to consider integrals over subsets of such sets Ω if we do not have an appropriate way of characterizing the *size* of these subsets. Indeed, whenever we think about integrating, we think of computing the area "under" some function. However, when we say "under" a function, we are making a statement about the sizes of subsets of the function's domain. Indeed, every integral of a function has an associated *integral measure*.

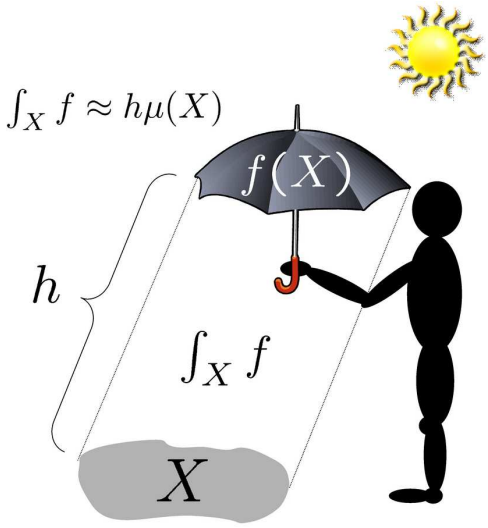

Figure 2: *This figure illustrates, conceptually, the necessity of having a measure μ on subsets A of the domain of a function f we want to integrate. Indeed, the approximation of the volume of shade ($\int_X f$) is the area of the shade spot on the ground ($\mu(X)$), times the "height" of the umbrella (h).*

Figuratively, we can think about these ideas by imagining we are outside with an umbrella on a sunny day. We might ask: How much shade is my umbrella providing? If we think about this in terms of the total volume of shade (i.e. the "integral under the umbrella"), we must not only consider the area of the umbrella (the range of our function), but also the *size* of the dark spot our umbrella makes on the ground. This is illustrated in Fig. 2. Another way of looking at it is realizing that in order to compute the integral $\int_X f$, we need a way to characterize the size of the set X in the left-most graph of Fig. 3. Therefore, in the discussion that follows, when we say "measurable function" f, we will always assume that we are talking about measurable functions $f : \Omega \to \bar{\mathbb{R}}$, where Ω is now part of the *measure space* $(\Omega, \mathcal{F}, \mu)$. Indeed, we are essentially combining our discussions of measures and measurable functions.

Let us first consider the Riemann integral of functions that map real numbers to real numbers. Riemann integration should be familiar to the reader from elementary real analysis. For the reader's convenience, we will define the Riemann integral using a definition from an introductory analysis textbook. Schumacher writes on pages 214 and 215 the following definitions [Sc].

Definition 3.1.1. Let a and b be real numbers with $a < b$. Then any set of the form $P = \{x_0, x_1, \ldots, x_{n-1}, x_n\}$ that satisfies $a = x_0 < x_1 < \ldots < x_n =$ is called a *partition* of $[a, b]$. The intervals $[x_{i-1}, x_i]$ are called the *subintervals* of $[a, b]$ determined by P. The *mesh* of the partition P is the length of its longest subinterval. Then, if f is a real-valued function whose domain contains the interval $[a, b]$, a *Riemann sum* for f corresponding to a partition P is

$$\mathcal{R}(f, P) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}),$$

where $x_i - 1 \leq x_i^* \leq x_i$ for each $i \leq n$.

Now, notice the kind of mathematical formalism that is being developed here for the Riemann integral. We begin by partitioning the domain of the function $f : \mathbb{R} \to \mathbb{R}$ into intervals. Think about an analogous procedure for a measurable function. How can we break up the set Ω? This set is completely general; it certainly does not have to be the set of real numbers. So, how do you break up, for example, the set of words in this paper into intervals? It is obvious that such a feat is intractably difficult, if not impossible. We can already see that the Riemann integral might not be the best approach. However, let us see what the actual definition is for this kind of integral. Again, from Schumacher:

Definition 3.1.2. Let $a, b \in \mathbb{R}$ with $a < b$. Let f be a real-valued function whose domain contains the interval $[a, b]$. We say that f is *Riemann Integrable* on $[a, b]$ if there exists a real number I such that for all $\epsilon > 0$ there exists $\delta > 0$ such that $|\mathcal{R}(f, P) - I| < \epsilon$ whenever $\mathcal{R}(f, P)$ is a Riemann sum for f coresponding to a partition of $[a, b]$ of mesh less than δ. We denote I by $\int_a^b f$ and call it *The Riemann Integral* of f over $[a, b]$.

Consider that what we are doing here is that we are considering a real-valued function f on an interval $[a, b]$. Then, we are breaking up $[a, b]$ into small subintervals. By making small rectangles out of these subintervals, we are able to compute the approximate area under the function by adding together the areas of the rectangles. We say that the integral of the function is the value of this

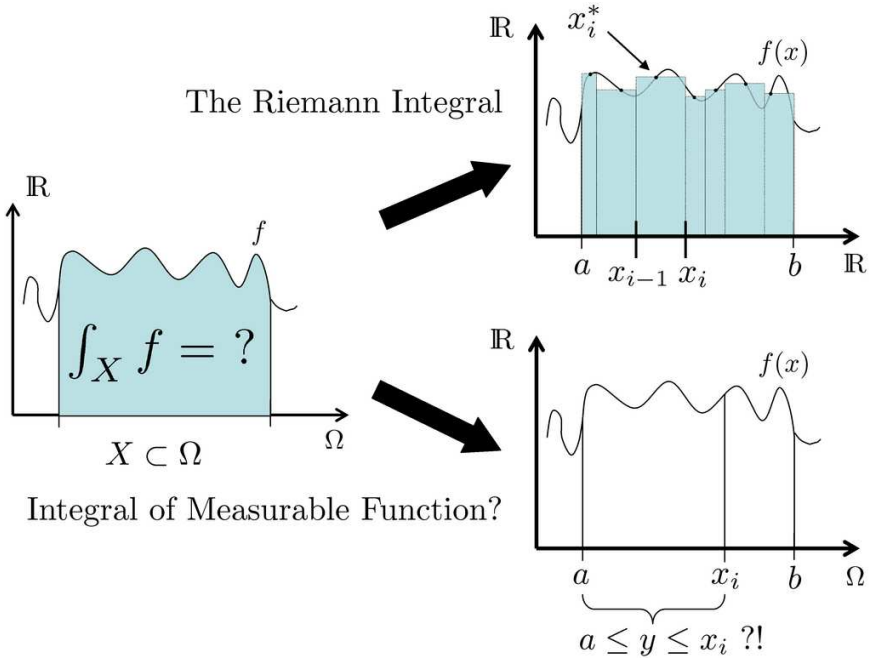

Figure 3: *Here we have an illustration of the terms in the definition of the Riemann integral. We also illustrate the difficulty of assigning similar definitions to arbitrary measurable functions.*

approximation as the partitioning becomes arbitrarily fine. The relevant terms in Def. 3.1.1 and Def. 3.1.2 are shown in Fig. 3. So, what can we learn from Riemann integrals that we can apply to our measurable functions? First of all, we certainly cannot partition our domain Ω into finer and finer pieces. As shown in Fig. 3, we may not even have greater than and less than relations on the set Ω. Thus, a partitioning of the set Ω into intervals is not plausible. However, the idea of adding up simple areas, such as rectangles, seems promising. Indeed, if we can somehow approximate our measurable function by a set of simple objects, then we might be able to compute an integral based on a finer and finer approximation of the measurable function, just as we did in the Riemann integral case. This is the main idea of the Lebesgue integral. Indeed, as we shall see, instead of breaking up our domain into finer and finer pieces to create *vertical* rectangles, it will be more advantageous to break up the *range* of our measurable functions and add up *horizontal* strips. However, we can't exactly add up horizontal strips, so what we really mean is that we add up the integrals of functions that approximate our function on these horizontal strips. This is a little difficult to think about right now without any explicit mathematics, however, we will attempt to illustrate this idea with a picture

(Fig. 4). The figure we have is a little teaser for the discussion that will follow, as we introduce some notions that seem a little foreign right now, such as the concept of "step functions". However, it will all be clearer by the end of the next section.

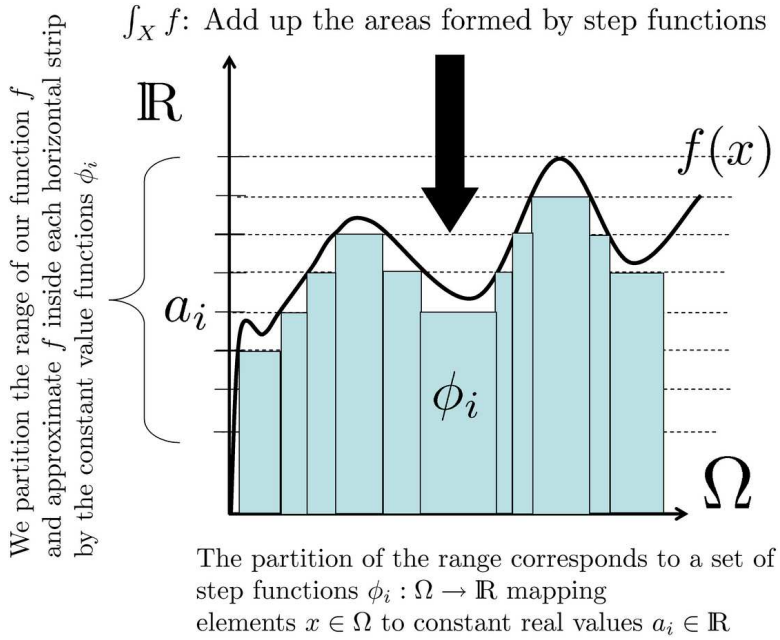

Figure 4: *This illustrates a way we can approximate the integral of an arbitrary measurable function. The idea is based on partitioning the range of the function and approximating each horizontal strip with step functions.*

3.2 Simple Functions

Motivated by the previous discussion, we make the following definitions (following Athreya on page 49)

Definition 3.2.1. Let $A \in \mathcal{F}$. We define a function $I_A : \Omega \to \mathbb{R}$ by

$$I_A(a) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{if } a \notin A \end{cases}$$

We call this function I_A the *characteristic function* of A.

Definition 3.2.2. A function $\phi : \Omega \to \bar{\mathbb{R}}$ is called *simple* if there exists a finite set of elements $\{a_1, a_2, \ldots, a_n\} \subset \bar{\mathbb{R}}$ and mutually disjoint measurable sets $A_1, A_2, \ldots, A_n \in \mathcal{F}$ such that

$$\phi = \sum_{i=1}^{n} a_i \, I_{A_i}.$$

Let us now show that the simple functions defined above are measurable using Def. 2.3.2. This will give us a better idea of how these definitions are implemented.

Theorem 3.2.3. *Simple functions are measurable.*

Proof. Suppose we have a simple function $\phi : \Omega \to \bar{\mathbb{R}}$ on $(\Omega, \mathcal{F}, \mu)$. Then, there exists a set of real numbers $\{a_1, a_2, \ldots, a_n\}$ and mutually disjoint measurable sets $A_1, A_2, \ldots, A_n \in \mathcal{F}$ such that

$$\phi = \sum_{i=1}^{n} a_i \, I_{A_i}.$$

Consider some point $y \in \Omega$. If $y \notin A_i$ for all $1 \leq i \leq n$, then $\phi(y) = 0$ by Def. 3.2.1. Next, suppose $y \in A_j$ for some $1 \leq j \leq n$. Then, since our sets A_i $(1 \leq i \leq n)$ are mutually disjoint, $I_{A_j}(y) = 1$ and $I_{A_i}(y) = 0$ for all $i \neq j$. Therefore, $\phi(y) = a_j$ (notice that this must be true for all $y \in A_j$). Since our $y \in \Omega$ was arbitrary, we find that ϕ can only take on the values $0, a_1, a_2, \ldots, a_n \in \bar{\mathbb{R}}$. Now, choose an arbitrary point $x \in \bar{\mathbb{R}}$ and consider the interval $(-\infty, x)$. From our considerations of ϕ above, and from the fact that $a_i \in (-\infty, x)$ whenever $a_i < x$, the inverse image of ϕ on this interval must be the union of all the A_i's for which $a_i < x$. Also, if $x > 0$, then $0 \in (-\infty, x)$ and we include the set $(\cup_{i=1}^{n} A_i)^{\mathcal{C}}$ in the union of the A_i's to construct the inverse image of ϕ on $(-\infty, x)$. However, $\mathcal{F}$ contains all of the sets A_i and is closed under complementation and finite unions. Thus, $\mathcal{F}$ always contains our inverse image $(-\infty, x)$ for all $x \in \bar{\mathbb{R}}$. So, the theorem follows from Def. 2.3.2. $\square$

As we shall see, these simple functions are what we will use to approximate the integral using the method outlined in Fig.4. In the figure, however, we referred to these functions as step-functions. This is not necessarily true, as we now know from the definition. However, step functions certainly *are* simple functions because they take on a finite set of values by definition. It's just that the converse is not necessarily true. Anyway, these definitions provide us a way of approximating measurable functions and their integrals. Specifically, we will prove later that *any* measurable function can be written as a limit of a sequence of simple functions. However, before we prove this, let us define the integral of these simple functions. Indeed, this is analogous to defining the Riemann sum for the Riemann integral definition. Specifically, we will deal with positive valued simple functions, first. This is intuitively appealing because when we think of areas, we usually think of positive numbers, first.

Definition 3.2.4. Let $\phi : \Omega \to \bar{\mathbb{R}}$ be a nonnegative simple function. We can represent this function in the same way we represented it in Def. 3.2.2, except we require that $\{a_1, a_2, \ldots, a_n\} \subset \bar{\mathbb{R}}_+$. The *integral* of ϕ over Ω with respect to μ, denoted by $\int_\Omega \phi \, d\mu$, is defined as

$$\int_\Omega \phi \, d\mu \equiv \sum_{i=1}^{n} a_i \, \mu(A_i).$$

This definition makes intuitive sense because we are summing over the product of the function values a_i and the sizes of the corresponding sets A_i over which the function takes on the value a_i. Notice that the value of this integral must be greater than or equal to zero since the function μ is always positive and all of the values a_i are in $\bar{\mathbb{R}}_+$. We can also consider the integral of ϕ over some subset of $A \subset \Omega$. However, to make sure this subset is measurable, we require that $A \in \mathcal{F}$. Thus, we simply say that the integral of ϕ over the set $A \in \mathcal{F}$ is the integral over Ω of the simple function $I_A\phi$. This function must be simple because we know that for any sets A and B, $I_A I_B = I_{A\cap B}$. This follows directly from the definition of a characteristic function. Therefore, the function $I_A\phi$ is defined in the same way as in Def. 3.2.2, except we now substitute $A_i \cap A$ for A_i in the definition. Consequently, we say that $\int_A \phi \, d\mu \equiv \sum_{i=1}^{n} a_i \, \mu(A_i \cap A)$.

Before we do anything else, it is important to verify that our Def. 3.2.4 makes sense for *any* representation of the simple function f. Therefore, we have the following theorem, which I proved by completing Problem 2.17 on page 76 of Athreya's Book [At].

Theorem 3.2.5. *Let ϕ be a simple function on Ω. Suppose we have two different representations of*

ϕ given by

$$\phi = \sum_{i=1}^{n} a_i I_{A_i} \quad \textit{and} \quad \phi = \sum_{j=1}^{m} b_j I_{B_j},$$

where $A_i, B_j \in \mathcal{F}$ and $a_i, b_j \in \bar{\mathbb{R}}$ for all i, j. Then,

$$\int \phi \, d\mu = \sum_{i=1}^{n} a_i \mu(A_i) = \sum_{j=1}^{m} b_j \mu(B_j).$$

That is, the integral of ϕ is independent of its representation

Proof. First, consider that we know $\bigcup_{i=1}^{n} A_i = \bigcup_{j=1}^{m} B_j = \Omega$ from Def. 3.2.2. Therefore, it is clear that $A_i = \bigcup_{j=1}^{m} A_i \cap B_j$. Furthermore, recall that $\{A_i \mid 1 \leq i \leq n\}$ and $\{B_j \mid 1 \leq j \leq m\}$ are collections of mutually disjoint sets. Consequently, since sets commute under the intersection operation, the set $\{A_i \cap B_j \mid 1 \leq i \leq n, 1 \leq j \leq n\}$ must also be a collection of mutually disjoint sets. However, notice something interesting about these sets. First, suppose that $A_p \cap B_q = \emptyset$ for some p, q. Then, we know that $\mu(A_p \cap B_q) = 0$ from the definition of a measure. However, if $A_p \cap B_q \neq \emptyset$, then there exists some $y \in \Omega$ that is in both A_p and B_q. We also know that in this case, by the mutual disjointness of these sets, $y \notin A_i$ for all $i \neq p$ and $y \notin B_j$ for all $j \neq q$. Therefore, if we consider $\phi(y)$, we realize from our definition of a characteristic function that $\phi(y) = a_p I_{A_p}(y) = a_p = b_q I_{A_q}(y) = b_q$. In other words, whenever $\mu(A_i \cap B_j) \neq 0$, we know that $a_i = b_j$. We will keep this in mind as we conclude our argument. From the properties of the measure μ given in Def. 2.1.5, we find that

$$\begin{aligned}
\sum_{i=1}^{n} a_i \mu(A_i) &= \sum_{i=1}^{n} a_i \mu\left(\bigcup_{j=1}^{m} A_i \cap B_j\right) = \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \mu(A_i \cap B_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} b_j \mu(B_j \cap A_i) = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} \mu(B_j \cap A_i) \\
&= \sum_{j=1}^{m} b_j \mu\left(\bigcup_{i=1}^{n} B_j \cap A_i\right) = \sum_{j=1}^{m} b_j \mu(B_j).
\end{aligned}$$

We now conclude that our definition for the integral of a positive simple function is well defined. □

Now that we have established the definition, it is important to derive a few properties of these integrals of simple functions. This will allow us to later prove similar properties of integrals of measurable functions. The proof that follows is my own for parts *i.* and *ii.*. The proof for part *iii.* is outlined on page 117 in Bogachev's book [Bo].

Theorem 3.2.6. *Let $\phi, \psi : \Omega \to \bar{\mathbb{R}}_+$ be two nonnegative simple functions. Also, let $X \in \mathcal{F}$. Since our integrals are independent of the simple function representation (by Thm. 3.2.5), we will suppose that $\{a_1, a_2, \ldots, a_n\} \subset \bar{\mathbb{R}}_+$ and $\{A_1, A_2, \ldots, A_n\} \subset \mathcal{F}$ are the set of distinct values and collection of mutually disjoint subsets of Ω, respectively, that represent ϕ. Similarly, $\{b_1, b_2, \ldots, b_m\} \subset \bar{\mathbb{R}}_+$ and $\{B_1, B_2, \ldots, B_m\} \subset \mathcal{F}$ will be the sets for ψ. We can do this by simply finding the finite set of distinct values ($\{a_i\}$ and $\{b_j\}$) that the functions ϕ and ψ evaluate to on their domain, and then constructing the sets A_i and B_j by finding $A_i = \phi^{-1}(a_i)$ and $B_j = \psi^{-1}(b_i)$. Then,*

i. $\int_X (\alpha\phi)\, d\mu = \alpha \int_X \phi\, d\mu$ *for* $\alpha \in \mathbb{R}$ *such that* $\alpha \geq 0$

ii. $\int_X (\phi + \psi)\, d\mu = \int_X \phi\, d\mu + \int_A \psi\, d\mu$.

iii. If $\phi(x) \leq \psi(x)$ *almost everywhere in* X, *then* $\int_X \phi\, d\mu \leq \int_X \psi\, d\mu$. *We call this property the monotinicity property.*

iv. If $\phi(x) = \psi(x)$ *almost everywhere in* X, *then* $\int_X \phi\, d\mu \leq \int_X \psi\, d\mu$.

Proof.
($i.$): This part follows directly from the definition. We see that

$$\int_X (\alpha\phi)\, d\mu = \sum_i^n \alpha a_i \mu(A_i \cap X) = \alpha \sum_i^n a_i \mu(A_i \cap X) = \alpha \int_X \phi\, d\mu$$

($ii.$): Again, just as in the proof of Thm. 3.2.5, we realize from the definition of simple function that $\{A_i \cap B_j \,|\, 1 \leq i \leq n, 1 \leq j \leq m\}$ is a collection of mutually disjoint sets that cover Ω. This means that for all $y \in \Omega$, $\sum_{i=1}^n I_{A_i}(y) = \sum_{j=1}^m I_{B_j}(y) = 1$. Also, as before, if $A_p \cap B_q \neq \emptyset$ for some p, q, then we know that for all $x \in A_p \cap B_q$, $\phi(x) + \psi(x) = a_p + b_q$. We may now compute our integral.

Specifically,

$$\begin{aligned}\int_X (\phi+\psi)\,d\mu &= \int_X \left(\sum_{i=1}^n a_i I_{A_i} + \sum_{j=1}^m b_j I_{B_j}\right) d\mu = \int_X \left(\sum_{i=1}^n a_i I_{A_i} \sum_{j=1}^m I_{B_j} + \sum_{j=1}^m b_j I_{B_j} \sum_{i=1}^n I_{A_i}\right) d\mu \\ &= \int_X \left(\sum_{i=1}^n \sum_{j=1}^m a_i I_{A_i\cap B_j} + \sum_{i=1}^n \sum_{j=1}^m b_j I_{B_j\cap A_i}\right) d\mu = \sum_{i=1}^n \sum_{j=1}^m (a_i+b_j)\mu(A_i\cap B_j\cap X) \\ &= \sum_{i=1}^n a_i \sum_{j=1}^m \mu(A_i\cap B_j\cap X) + \sum_{j=1}^m b_j \sum_{i=1}^n \mu(A_i\cap B_j\cap X) \\ &= \sum_{i=1}^n a_i\mu(A_i\cap X) + \sum_{j=1}^m b_i\mu(B_j\cap X) = \int_X \phi\,d\mu + \int_X \psi\,d\mu.\end{aligned}$$

(*iii.*): Consider the set $\Lambda = \{x \in X \mid \phi(x) \leq \psi(x)\}$. Since the relation specified in the set holds almost everywhere, we know that $\mu(X \setminus \Lambda) = 0$. Also, $\Lambda \in \mathcal{F}$ since the functions ϕ and ψ are measurable. Suppose that $c = \sup_{x\in X}\{|\phi(x)| + |\psi(x)|\}$. Then, since $\phi > \psi$ only on the set $X \setminus \Lambda$, then we know that for all $x \in X$, $\psi(x) - \phi(x) + cI_{X\setminus\Lambda}(x) \geq 0$. Notice that the term $cI_{X\setminus\Lambda}$ is a simple function! This is great! We have a sum of simple functions that is non-negative. Therefore, this sum is itself a simple function. Thus, we know that the integral of this sum will also be greater than or equal to zero, just by our definition of the integral. We can now use our previous results, part (*ii.*) and (*i.*), to conclude that

$$\begin{aligned}&\int_X (\psi - \phi + cI_{X\setminus\Lambda})\,d\mu = \int_X \psi\,d\mu - \int_X \phi\,d\mu + c\int_X I_{X\setminus\Lambda}\,d\mu = \int_X \psi\,d\mu - \int_X \phi\,d\mu + c\mu(X\setminus\Lambda) \geq 0 \\ &\int_X \psi\,d\mu - \int_X \phi\,d\mu \geq 0 \quad \Rightarrow \quad \int_X \phi\,d\mu \leq \int_X \psi\,d\mu\end{aligned}$$

(*iv.*): It is clear that $\phi = \psi$ almost everywhere implies that $\phi \leq \psi$ and $\phi \geq \psi$ for every $x \in X$ such that $\phi(x) = \psi(x)$. Indeed, by part (*iii.*), this implies that $\int_X \phi\,d\mu \leq \int_X \psi\,d\mu$ and $\int_X \phi\,d\mu \geq \int_X \psi\,d\mu$. We may now conclude that $\int_X \phi\,d\mu = \int_X \psi\,d\mu$ □

We now have some very important results concerning the integrals of simple functions. Specifically, we are now able to add and subtract simple functions from each other and be confident that the integrals add and subtract, also. We may also pull constants out of integrals with confidence. Finally, parts (*iii.*) and (*iv.*) of Thm. 3.2.6 tell us that sets of measure zero have no effect on the integral. Conceptually, we can think about this in terms of our picture on Fig. 2. If we have a shade spot

with no area ($\mu(X) = 0$), then our volume of shade is also zero. Indeed, if there is no shade spot ($X = \emptyset$), or if the shade spot is grainy (e.g. a Cantor set or a countable subset of the real numbers), our umbrella provides us with no shade at all.

We finish this discussion by connecting simple functions to measurable functions. This is necessary so that we can extend our definition of the integral to arbitrary measurable functions. The following theorem is a very powerful and crucial result. It will basically tell us that *any* measurable function is a limiting function of a sequence of simple functions. The proof given here follows the one presented on page 62 in Adams' book [Ad].

Theorem 3.2.7. *Let $f : \Omega \to \bar{\mathbb{R}}_+$ be a nonnegative measurable function. There exists a sequence of nonnegative simple functions*

$$0 \leq \phi_1 \leq \phi_2 \leq \phi_3 \leq \ldots \leq f$$

such that $\phi_i \to f$ pointwise. Using our special notation, we say that $\phi_i \uparrow f$. Also, if f is bounded, then $\phi_i \to f$ uniformly.

Proof. We will first define the sequence (ϕ_i). Let us pick some arbitrary index $n \in \mathbb{N}$. We will consider the interval $[0, n)$ in $\bar{\mathbb{R}}$. Let us break up this interval into smaller subintervals of length $1/2^n$. Thus, we will have $n/(1/2^n) = n2^n$ different subintervals $\mathcal{I}_j$. We can explicitly write them as

$$\mathcal{I}_j = \left\{ y \in \bar{\mathbb{R}} \mid \frac{j-1}{2^n} \leq y < \frac{j}{2^n} \right\} \text{ for } 1 \leq j \leq n2^n.$$

Notice that these subintervals simply divide up the interval $[0, n)$ into $n2^n$ equally sized pieces. However, when we move from n to $n+1$, we increase the number of pieces to $(n+1)2^{n+1}$. From our definition, we see that this is done by cutting in half all of the pieces in the n^{th} case and adding 2^{n+1} of these pieces to the end of the $[0, n)$ interval so that our new half-sized pieces now cover the interval $[0, n+1)$. This is illustrated in Fig. 5. We will now consider the sets $A_j = f^{-1}(\mathcal{I}_j)$ and $B_n = f^{-1}([n, \infty])$. Notice that the sets $\mathcal{I}_1, \mathcal{I}_2, \ldots \mathcal{I}_{n2^n}$, and $[n, \infty]$ cover the entire $\bar{\mathbb{R}}_+$ set. Therefore, the collection $\{A_1, A_2, \ldots, A_{n2^n}, B_n\}$ forms a cover of Ω, since our function f is well-defined. We will now define our simple function ϕ_n as

$$\phi_n(x) = \sum_{j=1}^{n2^n} \left(\frac{j-1}{2^n} \right) I_{A_j} + nI_{B_n}$$

This definition allows ϕ_n to take on a particular value for the different regions in the partition of our range of f. Indeed, if we look at Fig. 5, we can see that ϕ_n simply approximates the variation

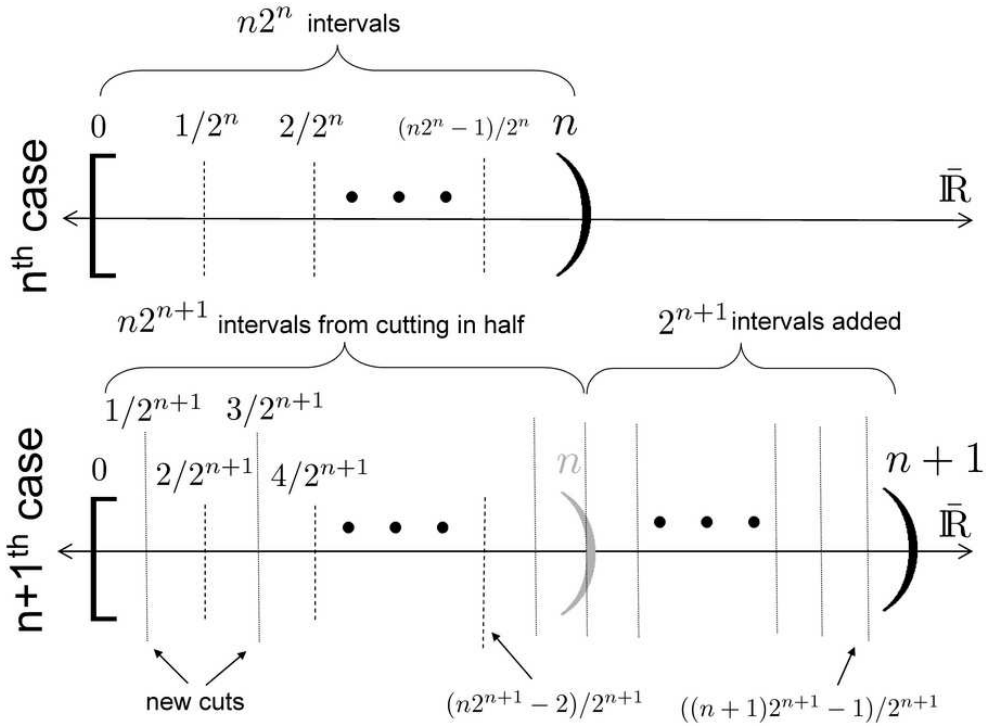

Figure 5: *This illustrates the procedure of partitioning our interval* $[0, n)$ *into subintervals. We also show how this partition changes as* $n \to n+1$.

in the range in f over a particular interval $(\mathcal{I}_j)$, by a constant value $\frac{j-1}{2^n}$ if we are looking at some point $x \in \Omega$ in the domain such that $f(x) < n$. But, if $f(x) \geq n$, then we allow $\phi_n(x)$ to just be n. So, what are we doing here conceptually? We are approximating our function f by partitioning its range into horizontal "strips". Indeed, as we mentioned previously, this is just like the technique used for the Riemann integral, except instead of partitioning the domain, we are partitioning the range. It should be clear that as $n \to \infty$, our partition of the range becomes arbitrarily fine and in each case covers the entire range of f (all of $\bar{\mathbb{R}}_+$). Therefore, as $n \to \infty$, we expect our sequence (ϕ_n) to converge to f. We shall see that this is the case in the last steps of the proof.

So, motivated by the previous discussion and figures, consider that on any set A_j, our function f has to satisfy, for each $x \in A_j$,

$$\frac{j-1}{2^n} \leq f(x) < \frac{j}{2^n} \text{ and } \phi_n(x) = \frac{j-1}{2^n}.$$

Consequently, $\phi_n(x) \leq f(x)$ for all $x \in A_j$. This must be true for any $1 \leq j \leq n2^n$. Also, for any

given B_n, we have that $n \leq f$ and $\phi_n = n$ on this set. This means that $\phi_n(x) \leq f(x)$ for all $x \in B_n$. Once again, since our collection of A_i's with B_n cover Ω, $\phi_n(x) \leq f(x)$ for all $x \in \Omega$. We now want to show that $\phi_n \leq \phi_{n+1}$ for all $n \in \mathbb{N}$. Thus, consider one our subintervals $\mathcal{I}_j = [(j-1)/2^n, j/2^n)$. We can cut this interval in half. This is illustrated in Fig. 6. From Fig. 6, we see that $\mathcal{I}_j = \mathcal{I}'_j \cup \mathcal{I}''_j$. We

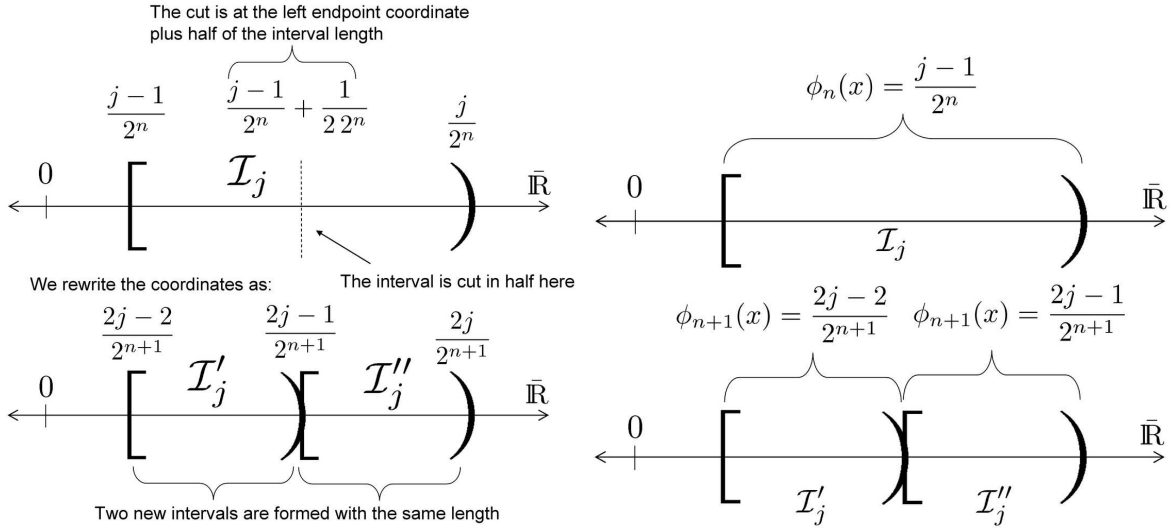

Figure 6: *This illustrates the procedure of cutting our interval $\mathcal{I}_j$ in half. Two new intervals $\mathcal{I}'_j$ and $\mathcal{I}''_j$ are formed. The values of our simple functions ϕ_n on these intervals are also shown in the figures on the right.*

now consider the sets $A_j = f^{-1}(\mathcal{I}_j)$, $A'_j = f^{-1}(\mathcal{I}'_j)$, and $A''_j = f^{-1}(\mathcal{I}''_j)$. From the way we constructed our simple functions ϕ_n, and from the comparison of Fig. 5 and Fig. 6, it should be clear that $\phi_n(x) = (j-1)/2^n$ for all $x \in A_j$, $\phi_{n+1}(x) = (j-1)/2^n$ for all $x \in A'_j$, and $\phi_{n+1}(x) = (2j-1)/2^{n+1}$ for all $x \in A''_j$. Also, $A_j = A'_j \cup A''_j$. Now consider any $x \in A_j$. Then, either $x \in A'_j$ or $x \in A''_j$. In the first case,

$$\phi_n(x) = \frac{j-1}{2^n} = \phi_{n+1}(x).$$

In the second case,

$$\phi_n(x) = \frac{j-1}{2^n} = \frac{2j-2}{2^{n+1}} < \frac{2j-1}{2^{n+1}} = \phi_{n+1}(x).$$

We have now shown that $\phi_n \leq \phi_{n+1}$ for all $x \in A_j$. However, we started with an arbitrary A_j, so this must be true for all $1 \leq j \leq n2^n$. Finally, we also have that $\phi_n(x) \leq \phi_{n+1}(x)$ for all $x \in B_n$. This is because $B_{n+1} \subset B_n$ and if $x \in B_{n+1}$, then $\phi_n(x) = n < n+1 = \phi_{n+1}(x)$. Conversely, if

$x \notin B_{n+1}(x)$, then $\phi_{n+1}(x)$ has to be equal to the left-most endpoint of one of the added subintervals we show in Fig. 5. This, as seen in the figure, has to be greater than or equal to n. Therefore, $\phi_n(x) = n \leq \phi_{n+1}(x)$ for all $x \in B_n$ but not in B_{n+1}, as well. Therefore, since we already said that the sets A_j along with B_n form a cover of Ω, then $\phi_n \leq \phi_{n+1}$ on all of Ω!

What is left to do now is to show that our sequence (ϕ_n) converges pointwise to f. Therefore, consider an arbitrary $x \in \Omega$. Since we allowed f to map to the extended real numbers, let us first deal with the situation when $f(x) = \infty$. In this case, we know that $x \in B_n$ for all $n \in \mathbb{N}$, just by our definition of the sets B_n above. Therefore, $\phi_n(x) = n$ for all $n \in \mathbb{N}$. This is an unbounded increasing sequence and clearly $\phi_n(x) \to \infty$. We are now finished with this case.

Now we can suppose that $f(x)$ is some finite real number. This means that we can choose some $n_0 \in \mathbb{N}$ such that $f(x) < n_0$. Then, recall that for a given term in our sequence of simple functions ϕ_n, the definition of ϕ_n involves a partitioning of the interval $[0, n)$. Therefore, we know that for all $n > n_0$, our function value $f(x)$ must be located in one of the subintervals $\mathcal{I}_j$, since $f(x) < n_0$ implies $f(x) \in [0, n)$ for all $n > n_0$. So, choosing this particular $\mathcal{I}_j$, we know that

$$\frac{j-1}{2^n} \leq f(x) < \frac{j}{2^n}$$

for all $n > n_0$ from the definition of $\mathcal{I}_j$ for a particular n. However, from our definition of ϕ_n, we now know that $\phi_n(x) = (j-1)/2^n$. Therefore,

$$|f(x) - \phi_n(x)| = \left| f(x) - \frac{j-1}{2^n} \right| < \left| \frac{j}{2^n} - \frac{j-1}{2^n} \right| = \frac{1}{2^n}.$$

Since we can make $1/2^n$ arbitrarily small as we increase n, we have now proven that $\phi_n(x) \to f(x)$. Also, since our x was arbitrary, we have now shown the pointwise convergence. The last bit about uniform convergence should be easy to see. If f is bounded, then $f(x) < n_0$ for all $x \in \Omega$ and some $n_0 \in \mathbb{N}$. Therefore, from the preceding arguments,

$$|f(x) - \phi_n(x)| < \frac{1}{2^n} \quad \text{for all } x \in \Omega.$$

This is the condition for uniform convergence of $\phi_n \to f$. □

3.3 The Lebesgue Integral

After our discussion of simple functions, the reader should have some idea of how to define the integral of any measurable function. Specifically, it should be clear that our motivation is the approximation

of measurable functions by simple ones. Thus, we can begin right away with a *definition* of the integral of a nonnegative measurable function. In fact, mathematical literature tells us that there are two equivalent definitions for such an integral. We will present both and refer the reader to pages 51 and 52 in Athreya's book [Ath] for the outline of the proof that they are equivalent. The first definition is stated on page 61 in Adams' book [Ad], and the second definition is in [Ath] on page 50.

Definition 3.3.1. Let $f : \Omega \to \bar{\mathbb{R}}_+$ be a nonnegative measurable function and $X \in \mathcal{F}$. The integral of f over X with respect to μ is defined by

$$\int_X f\, d\mu = \sup \left\{ \int_X \phi\, d\mu \mid 0 \leq \phi \leq f \text{ where } \phi \text{ is a simple function.} \right\}.$$

Definition 3.3.2. Let $f : \Omega \to \bar{\mathbb{R}}_+$ be a nonnegative measurable function and $X \in \mathcal{F}$. The integral of f over X with respect to μ is defined by

$$\int_X f\, d\mu = \lim_{n\to\infty} \int_X \phi_n\, d\mu,$$

where $\{\phi_n\}_{n\geq 1}$ is any sequence of nonnegative simple functions such that $\phi_n(x) \uparrow f(x)$ for all $x \in X$.

We already know from our properties of integrals of simple functions, Thm. 3.2.6, that because $\phi_1 \leq \phi_2 \ldots$, then the sequence $(\int_X \phi_i\, d\mu)$ is also increasing. This means that our limit given in the right hand side of the equality in Def. 3.3.2 is well defined. However, it remains to be seen that the definition of the integral in Def. 3.3.2 remains the same for any given sequence of simple functions such that $\phi_n \uparrow f$. So, we will now show this in a theorem. The proof follows the one outlined on page 50 in Athreya's book [Ath].

Theorem 3.3.3. *Let $\{\phi_n : \Omega \to \bar{\mathbb{R}}_+\}_{n\geq 1}$ and $\{\psi_n : \Omega \to \bar{\mathbb{R}}_+\}_{n\geq 1}$ be two sequences of simple nonnegative functions such that $\phi_n(x) \uparrow f(x)$ and $\psi_n(x) \uparrow f(x)$ for all $x \in X$ where $X \in \mathcal{F}$. Then,*

$$\lim_{n\to\infty} \int_X \phi_n\, d\mu = \lim_{n\to\infty} \int_X \psi_n\, d\mu.$$

Proof. First we will consider a particular function ψ_N where $N \in \mathbb{N}$. We want to show that for any real number $\rho \in \mathbb{R}$ such that $0 < \rho < 1$, we have that

$$\lim_{n\to\infty} \int_X \phi_n\, d\mu \geq \rho \int_X \psi_N\, d\mu.$$

So, in order to do this, we will allow our simple function ψ_N be represented by $\psi_N = \sum_{i=1}^{p} a_i I_{A_i}$. Then, we will consider the set $D_n = \{x \in X \mid \phi_n(x) \geq \rho\psi_N(x)\}$. Now, suppose that we have some $y \in D_n$. Then, since this means that $\phi_n(y) \geq \rho\psi_N(y)$. However, since $\phi_n(x) \uparrow f(x)$ for all $x \in X$, then $\phi_{n+1}(y) \geq \phi_n(y) \geq \rho\psi_N(y)$, as well. We conclude that $y \in D_{n+1}$ and so, by a trivial induction, $D_1 \subseteq D_2 \subseteq \ldots$. Now consider the set $D = \{x \in X \mid f(x) \geq \rho\psi_N(x)\}$. We want to show that $D = \bigcup_{n=1}^{\infty} D_n$.
($\subseteq$): Suppose that $y \in D$. This means that $f(y) \geq \rho\psi_N(y)$. However, by our construction of the (ψ_i) sequence, we know that actually, $f(y) \geq \psi_N(y) > \rho\psi_N(y)$. Now, since $\phi_n(y) \uparrow f(y)$, we can choose an $M \in \mathbb{N}$ such that (since $f(y) - \rho\psi_N(y) > 0$

$$f(y) - \phi_M(y) < f(y) - \rho\psi_N(y) \quad \Rightarrow \quad \phi_M(y) > \rho\psi_N(y).$$

We conclude that $y \in D_M$ and thus $y \in \bigcup_{n=1}^{\infty} D_n$.
($\supseteq$): Let $y \in \bigcup_{n=1}^{\infty} D_n$. This means that there is some $M \in \mathbb{N}$ such that $f(y) \geq \phi_M(y) \geq \rho\psi_N(y)$. We now see that $y \in D$ and we are done. Now that we have shown $D_n \uparrow D$, consider again that $\psi_N(y) \leq f(y)$ for all $y \in X$. Therefore, $D = X$. This means that for any particular D_n, $D_n \cup D_n^{\mathcal{C}} = X$. Therefore, we can write our simple functions ϕ_n as $\phi_n = \phi_n I_{D_n} + \phi_n I_{D_n^{\mathcal{C}}}$. From our properties of integrals of simple functions,

$$\int_X \phi_n \, d\mu \geq \int_X \phi_n I_{D_n} \geq \rho \int_X \psi_N I_{D_n} \, d\mu = \rho \sum_{i=1}^{p} a_i \mu(A_i \cap D_n \cap X).$$

It should now be clear that since $D_n \uparrow X$, then as $n \to \infty$, we have that $\mu(A_i \cap D_n \cap X) \uparrow \mu(A_i \cap X)$. Thus, we conclude that

$$\lim_{n\to\infty} \int_X \phi_n \, d\mu \geq \rho \lim_{n\to\infty} \sum_{i=1}^{p} a_i \mu(A_i \cap D_n \cap X) = \rho \int_X \psi_N \, d\mu.$$

However, we can let ρ be arbitrarily close to 1, so we find that $\lim_{n\to\infty} \int_X \phi_n \, d\mu \geq \int_X \psi_N \, d\mu$ for every $N \in \mathbb{N}$. Therefore, we now see that

$$\lim_{n\to\infty} \int_X \phi_n \, d\mu \geq \lim_{n\to\infty} \int_X \psi_n \, d\mu.$$

The theorem follows by a parallel proof showing that we can get the opposite inequality to the one in the equation above. Indeed, there is no reason we cannot reverse our argument by simple switching all the inequalities. So, we are now done. $\square$

Now that we have a well-defined notion of the integral of a measurable function, we can derive some properties of these Lebesgue integrals. They all naturally follow from the properties given for the integrals of simple functions. Indeed, using our Def. 3.3.1 and Def. 3.3.2, we can treat each Lebesgue integral in terms of integrals of simple functions. Therefore, all of the arguments we used in the previous section apply to this one. This is a great thing! We have managed to tackle the integrals of very abstract, general functions through approximation by simple functions. It is worthwhile to think about this in contrast to Riemann integration with which we are familiar with. This new Lebesgue method is much more powerful because it allows us to integrate functions that do not have a very structured domain. Indeed, our integral over $X \in \mathcal{F}$ only requires that X have an appropriate measure μ and is part of the σ-algebra $\mathcal{F}$. This means that X can be any of the various examples we mentioned in Sec. 2.2.

We will now present the following results concerning the behavior of Lebesgue integrals. These properties are *identical* to the ones for integrals of simple functions. Indeed, it should be clear that all of the following follows from Thm. 3.2.6, Def. 3.3.1, and Def. 3.3.2. The statements of these theorems are on page 49 in Athreya [Ath]. The proofs, which will not be included here, are outlined on pages 189 to 191 in [Bu].

Theorem 3.3.4. *Let f and g be two measurable nonnegative functions on $(\Omega, \mathcal{F}, \mu)$. Then, if $X \in \mathcal{F}$,*

i. For $\alpha \geq 0$ and $\beta \geq 0$, $\int_X (\alpha f + \beta g)\, d\mu = \alpha \int_X f\, d\mu + \beta \int_X g\, d\mu$

ii. If $f \geq g$ almost everywhere in X, then $\int_X f\, d\mu \geq \int_X g\, d\mu$.

iii. If $f = g$ almost everywhere in X, then $\int_X f\, d\mu = \int_X g\, d\mu$.

The reader may have noticed that we haven't dropped the qualifier "nonnegative" when describing the integrals of our measurable functions. Indeed, all of our theorems up to this point have involved positive valued functions. However, we know that we can take integrals of functions that yield negative values, as well. Indeed, we may recall that the Riemann integral definition made no requirements concerning the values of f. If f was negative at a particular sampling point x_i^*, then we simply subtracted the area of the rectangle $f(x_i^*)(x_i - x_{i-1}) = -|f(x_i^*)(x_i - x_{i-1})| < 0$ in our Riemann sum. Thus, we want to define the integral of a measurable function *in general* (regardless of whether it is positive or negative). The following definition is listed on page 54 in Athreya's book.

Definition 3.3.5. Let f be a real valued measurable function on $(\Omega, \mathcal{F}, \mu)$. Let $f^+ \equiv fI_{\{f\geq 0\}}$ and $f^- \equiv -fI_{\{f<0\}}$. The integral of f with respect to μ over $X \in \mathcal{F}$ is defined as

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

provided that at least one of the integrals on the right side is finite.

Now, notice that both f^+ and f^- as they are defined above are nonnegative measurable functions. Thus, the definition makes sense *provided* either the f^+ integral or f^- integral is finite. Now, we repeat this proviso because it is important. Recall from our discussion of the extended real numbers that we cannot perform the operations $\infty - \infty$ or $-\infty + \infty$. Getting the latter computations in an integral would be disastrous because the results of these operations are not defined. This is why we need the proviso given.

Before we move on, it is important to mention a few more definitions. Indeed, mathematicians often refer to seemingly archaic things like $\mathcal{L}^p$ spaces, $\mathcal{L}$ integrability, etc. So, let us make these things clear. We will use the language in Arthreya's book (page 54-55) [Ath].

Definition 3.3.6. Let f be a real valued measurable function on $(\Omega, \mathcal{F}, \mu)$. We say that f is *integrable*, or *$\mathcal{L}$-integrable*, with respect to μ if $\int_\Omega |f|\, d\mu < \infty$. Notice that since $|f| = f^+ + f^-$, then this condition is the same as saying $\int_\Omega f^+\, d\mu < \infty$ and $\int_\Omega f^-\, d\mu < \infty$. Whenever f is integrable, we also say that $f \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$.

Definition 3.3.7. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space that $0 < p \leq \infty$. Then, the collection of functions $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is defined as

$$\mathcal{L}^p(\Omega, \mathcal{F}, \mu) = \{f \,:\, |f|^p \text{ is integrable with respect to } \mu\} = \left\{ f \,:\, \int |f|^p\, d\mu < \infty \right\} \text{ for } 0 < p < \infty,$$

and

$$\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu) \equiv \left\{ \quad f \,:\, \mu(\{|f| > K\}) = 0 \text{ for some } K \in (0, \infty) \quad \right\}.$$

Notice that this last defined space, $\mathcal{L}^\infty$, is simply the set of all functions that are bounded on Ω except for possibly on a set $A \subset \Omega$ of measure zero. This is in accordance with our definition for the other spaces since we know the integrability of a measurable function is independent of subsets of measure zero of the domain of the function. With these definitions in mind, we see that it is trivial to show that the conditions of Thm. 3.3.4 hold for all functions $f \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$, as well.

3.4 Convergence Theorems

The last few sections provided us with the mathematical formalism for dealing with integrals of arbitrary measurable functions. We proved that these integrals are well-defined and possess certain natural properties. We have also shown that sets of measure zero do not influence the values of these integrals. Now what we want to do is to prove some fundamental theorems about these Lebesgue integrals. The first of these will be a fun, easy theorem that will be familiar to the reader from basic probability theory. Specifically, we already have the mathematical machinery to prove Markov's inequality! Although it might not look explicitly like the inequality we are familiar with, we shall see later that the following theorem is exactly what it claims to be. The proof is outlined on page 66 in Adams' book [Ad].

Theorem 3.4.1. *(Markov's inequality) Let f be a nonnegative measurable function. For $X \in \mathcal{F}$ and $\alpha > 0$, let $X_\alpha = \{x \in X \mid f(x) \geq \alpha\}$. Then,*

$$\mu(X_\alpha) \leq \frac{1}{\alpha} \int_X f \, d\mu.$$

Proof. We are given that $f(x) \geq \alpha$ on X_α. Therefore, by our properties of the Lebesgue integral,

$$\int_{X_\alpha} \alpha \, d\mu \leq \int_{X_\alpha} f \, d\mu.$$

However, our constant function α is a simple function and can be written as αI_Ω. But, because we are only integrating over the space X_α, we have that $\int_{X_\alpha} \alpha I_\Omega \, d\mu = \alpha\mu(\Omega \cap X_\alpha) = \alpha\mu(X_\alpha)$. Finally, since $X_\alpha \subseteq X$ and f is nonnegative, $\int_{X_\alpha} f \, d\mu \leq \int_X f \, d\mu$ (this follows easily from our definition of the Lebesgue integral in terms of approximating sequences of simple functions in Def. 3.3.1). So, we combine these inequalities to find that

$$\alpha\mu\left(X_\alpha\right) \leq \int_{X_\alpha} f \, d\mu \leq \int_X f \, d\mu.$$

We are now done. □

It is often the case that we only know about a sequence of measurable functions that converge to some other function. We begin to study the behavior of integrals with respect to such sequences by considering the first important convergence theorem in the theory of the integral. The proof that follows is outlined on pages 52 and 53 in Athreya's book [Ath].

Theorem 3.4.2. (*The monotone convergence theorem (MCT)*) *Let (f_n) be a sequence of nonnegative measurable functions on $(\Omega, \mathcal{F}, \mu)$ and let f be such a function, also. Suppose that $f_n \uparrow f$ almost everywhere on Ω. Then, if $X \in \mathcal{F}$,*

$$\int_X f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

Proof. We will take advantage of our approximation methods with simple functions by considering a sequence of simple functions $\{\phi_n\}_{n\geq 1}$ such that $\phi_n(x) \uparrow f(x)$ for all $x \in X$. Now, consider the set $A \in \mathcal{F}$ for which $f_n \uparrow f$. We know that $\mu(A^{\mathcal{C}}) = 0$, since the condition $f_n \uparrow f$ must hold almost everywhere. Now, consider some $N \in \mathbb{N}$ and a real number $0 < \rho < 1$. Also, let $D_n = \{x \in A \mid f_n(x) \geq \rho\phi_N(x)\}$. This should be familiar to the reader from the proof of Thm. 3.3.3. From the same arguments given in that proof, if $D = \{x \in A \mid f(x) \geq \rho\phi_N(x)\}$, then $D_n \uparrow D$. Similarly, since $\phi_N(x) \leq f(x)$ for all $x \in \Omega$ (by the construction of our approximating sequence of simple functions), then $D = A$. So, from the properties of the integral that we have established in the previous section,

$$\int_X f_n \, d\mu \geq \int_X f_n I_{D_n} \, d\mu \geq \rho \int_X \phi_N I_{D_n} \, d\mu \text{ for all } n \geq 1.$$

Once again, because $D_n \uparrow D$, and our measure function μ has the countable additivity property, we know that $\int_X \phi_N I_{D_n} \, d\mu \uparrow \int_X \phi_N I_A \, d\mu = \int_X \phi_N \, d\mu$, where $n \to \infty$. Consequently, we can take the following limits:

$$\lim_{n\to\infty} \int_X f_n \, d\mu \geq \rho \int_X \phi_N \, d\mu.$$

Then, since our $N \in \mathbb{N}$ was arbitrary, this must be true for all $N \in \mathbb{N}$. Finally, just as the Thm. 3.3.3, we let $\rho \to 1$. This allows us to conclude that $\lim_{n\to\infty} \int_X f_n \, d\mu \geq \int_X \phi_N \, d\mu$. Finally, by letting $N \to \infty$ and letting ϕ_N approach f, we get that

$$\lim_{n\to\infty} \int_X f_n \, d\mu \geq \int_X f \, d\mu.$$

Finally, since $f_n \uparrow f$ implies that $f_1 \leq f_2 \leq \ldots$, we see from our Thm. 3.3.4 that

$$\int_X f_n \, d\mu \leq \int_X f \, d\mu \text{ for all } n \in \mathbb{N}.$$

By taking the limit $n \to \infty$, we get the opposite inequality and, thus, the proof follows. □

This theorem allows us to make some great conclusions. One of the first things we can do is consider the behavior of summations and integrals. Often, especially in physics, we haphazardly exchange summation and integration symbols without every considering what the formal mathematical reason is behind the exchange. In the following corollary, we show that this makes sense for nonnegative measurable functions. The proof is on page 75 in Adams' book [Ad].

Corollary 3.4.3. *Let (f_n) be a sequence of nonnegative measurable functions on $(\Omega, \mathcal{F}, \mu)$. Then,*

$$\int_X \left(\sum_{i=1}^{\infty} f_i \right) = \sum_{n=1}^{\infty} \int_X f_n \, d\mu.$$

Proof. We just have to consider the functions $F_n = \sum_{i=1}^{n} f_i$. Then, it is clear, since the f_i's are nonnegative, $F_1 \leq F_2 \leq F_3 \leq \ldots$. This means $F_n \uparrow \sum_{i=1}^{\infty} f_i$. Also, all the F_n are measurable since they are sums of measurable functions. This means we can apply the MCT directly to the sequence (F_n) and the limiting function $\sum_{i=1}^{\infty} f_i$. So, by the MCT and the properties of the Lebesgue integral we get that for any $X \in \mathcal{F}$,

$$\int_X \left(\sum_{i=1}^{\infty} f_i \right) d\mu = \lim_{n\to\infty} \int_X F_n \, d\mu = \lim_{n\to\infty} \int_X \sum_{i=1}^{n} f_i \, d\mu = \lim_{n\to\infty} \sum_{i=1^n} \int_X f_i \, d\mu = \sum_{n=1}^{\infty} \int_X f_n \, d\mu.$$

□

The MCT also allows us to prove a theorem, or corollary, that is called *Fatou's lemma*. We mention all three qualifications of a provable statement because Fatou's lemma is the middle stepping stone between two major convergence theorems in integration theory. Indeed, we will need the result that follows to prove the famous Lebesgue dominated convergence theorem (DCT). Indeed, the primary purpose of this section is to prove the following sequence of theorems:

$$\text{MCT} \quad \Rightarrow \quad \text{Fatou's lemma} \quad \Rightarrow \quad \text{DCT}$$

Right now we are on this middle block. Let us state it and prove it. The proof will follow the two versions given in Adams on page 78 and Athreya on page 54.

Theorem 3.4.4. (*Fatou's lemma*) *Let (f_n) be a sequence of nonnegative measurable functions on $(\Omega, \mathcal{F}, \mu)$. Then,*

$$\liminf_{n\to\infty} \int_X f_n \, d\mu \geq \int_X \liminf_{n\to\infty} f_n \, d\mu.$$

Proof. Let us construct a new sequence of functions (g_n) such that for each $n \in \mathbb{N}$, $g_n(x) \equiv \inf\{f_j(x) \mid j \geq n\}$. Also suppose that we have a sequence of real numbers $\{a_n\}_{n\geq 1}$ defined by $a_n \equiv \inf\{\int_X f_j \, d\mu \mid j \geq n\}$. Then, from the definition of the infimum, we see that

$$g_1 \leq g_2 \leq \ldots \quad \text{and} \quad a_1 \leq a_2 \leq \ldots$$

However, from our definition of f and from the fact that the infimum limit, the supremum limit, and the limit of convergent sequences are all identical, we conclude that $\liminf_{n\to\infty} \int_X f_n \, d\mu = \lim_{n\to\infty} a_n$ and $\liminf_{n\to\infty} f_n = \lim_{n\to\infty} g_n$. However, because our $\inf\{f_j(x) \mid j \geq m\}$ is the *greatest lower bound* of the set of functions f_n where $n \geq m$, we know that for $n \geq m$, $f_n \geq g_m$. Hence, by the monotinicity property of the integral, $\int_X f_n \, d\mu \geq \int_X g_m \, d\mu$ for all $n \geq m$. Therefore, $a_m \geq \int_X g_m \, d\mu$ for all $m \in \mathbb{N}$ (since a_m is the *greatest* lower bound). Consequently, by the MCT,

$$\liminf_{n\to\infty} \int_X f_n \, d\mu = \lim_{n\to\infty} a_n \geq \lim_{n\to\infty} \int_X g_n \, d\mu = \int_X \liminf_{n\to\infty} f_n \, d\mu.$$

We are now done with the proof. □

We will now consider perhaps the most important result in integration theory. This is the primary result we will be using in proving later theorems. The proof of the theorem follows the one given on pages 78 and 79 in Adams' book [Ad].

Theorem 3.4.5. (*Lebesgue dominated convergence theorem (DCT)*) *Let (f_n) be a sequence of measurable functions on $(\Omega, \mathcal{F}, \mu)$ such that the sequence converges pointwise to some function f for every point $x \in \Omega$. Let $X \in \mathcal{F}$. Then, if there exists a function $g \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$ such that*

$$|f_n(x)| \leq g(x) \text{ almost everywhere in } X \text{ for every } n,$$

then the function f is also in $\mathcal{L}(\Omega, \mathcal{F}, \mu)$ and

$$\int_X f \, d\mu = \lim_{n\to\infty} \int_X f_n \, d\mu.$$

In addition,

$$\lim_{n\to\infty} \int_X |f - f_n| \, d\mu = 0.$$

Proof. We recognize right away by Fatou's lemma that

$$\int_X |f| \, d\mu = \int_X \liminf_{n\to\infty} |f_n| \, d\mu \leq \liminf_{n\to\infty} \int_X |f_n| \, d\mu \leq \int_X |g| \, d\mu < \infty,$$

since $g \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Thus, we conclude $f \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$, as well. To prove the rest of the theorem, consider our hypothesis that $|f_n| \leq g$ almost everywhere in X. Now, even though this is not true everywhere, we said in previous discussions that the integrals of two functions are the same if the two functions differ from each other only on a set of measure zero. This means that we can redefine the f_n in such a way that $|f_n| \leq g$ *everywhere* in X. From our previous discussions, this can be done without a loss of generality. Therefore, we conclude that $g + f_n$ is a nonnegative function for every $n \in \mathbb{N}$. We use Fatou's lemma again and find that

$$\int_X \liminf_{n\to\infty} (g + f_n)\, d\mu \leq \liminf_{n\to\infty} \int_X (g + f_n)\, d\mu.$$

But our function g does not change with n and so $\liminf_{n\to\infty}(g + f_n) = g + \liminf f_n = g + f$. Again, this last equality follows from the fact that the supremum limit, infimum limit, and the limit of convergence sequences are identical. This is something we have used over and over. Similarly, we can make the following statement about the integrals:

$$\liminf_{n\to\infty} \int_X (g + f_n)\, d\mu = \int_X g\, d\mu + \liminf_{n\to\infty} \int_X f_n\, d\mu.$$

Using the result we found from Fatou's lemma, we see that

$$\int_X f\, d\mu \leq \liminf_{n\to\infty} \int_X f_n\, d\mu.$$

Of course, we also have the case that $g - f_n$ is a nonnegative function. Thus, by a parallel argument to the one given above, if we replace all of the f_n's with $-f_n$ and f's with $-f$'s, we get that

$$-\int_X f\, d\mu \leq \liminf_{n\to\infty} \left(-\int_X f_n\, d\mu \right).$$

However, notice that for any sequence of real numbers (a_n), $\limsup_{n\to\infty}(-a_n) = -\limsup_{n\to\infty}(a_n)$. This is because we are simply reflecting our real numbers around the 0 pivot point. Indeed, we are simply switching all the inequalities, since $a < b$ implies $-a > -b$ for all $a, b \in \mathbb{R}$. Therefore, with this equality in mind, we see that

$$-\int_X f\, d\mu \leq -\limsup_{n\to\infty} \left(\int_X f_n\, d\mu \right).$$

Multiplying both sides of the inequality about by (-1) and flipping the inequality yields

$$\int_X f\, d\mu \geq \limsup_{n\to\infty} \left(\int_X f_n\, d\mu \right).$$

However, it is true in all cases that the infimum limit is less than or equal to the supremum limit of a sequence. Therefore, we have *squeezed* our integral $\int_X f\,d\mu$ so that it is now equal to

$$\int_X f\,d\mu = \liminf_{n\to\infty}\left(\int_X f_n\,d\mu\right) = \limsup_{n\to\infty}\left(\int_X f_n\,d\mu\right) = \lim_{n\to\infty}\left(\int_X f_n\,d\mu\right).$$

We have now proven the first part of the theorem. For the second part, instead of looking at the functions f_n and g, we look at $\tilde{f}_n \equiv |f - f_n|$ and $\tilde{g} \equiv g + |f|$. Again, we see that $\tilde{g} + \tilde{f}_n = g + |f| + |f - f_n| \geq g + |f| + ||f| - |f_n|| \geq g + |f| + |f| - |f_n|$ is a nonnegative function because $g + 2|f| \geq g \geq |f_n|$. In the $-$ case, $\tilde{g} - \tilde{f}_n = g + |f| - |f - f_n| \geq g + |f| - (|f| + |f_n|) = g - |f_n| \geq 0$ is a nonnegative function, also. Therefore, we can apply the exact same arguments we used about f_n and g. In this case, though, we have that f replaced by $\lim_{n\to\infty}|f - f_n| = |f - f| = 0$. Therefore, making the substitutions in the final step of the argument above, we get that

$$\int_X |f - f|\,d\mu = 0 = \lim_{n\to\infty}\int_X |f - f_n|\,d\mu.$$

We have now shown both parts of the theorem. □

An often used corollary of the DCT is something called the *bounded convergence theorem*, or BCT. It trivially follows from the previous theorem.

Corollary 3.4.6. (*The bounded convergence theorem (BCT)): Let $\mu(\Omega) < \infty$. Then, if there exists a real number $0 < k < \infty$ such that $|f_n| \leq k$ almost everywhere in Ω and $f_n \to f$ almost everywhere (for each $n \geq 1$), then*

$$\lim_{n\to\infty}\int_\Omega f_n\,d\mu = \int_\Omega f\,d\mu \text{ and } \lim_{n\to\infty}\int_\Omega |f_n - f|\,d\mu = 0.$$

Proof. Since our DCT was proven for any integrable function g such that $|f_n(x)| \leq g(x)$ almost everywhere in some $X \in \mathcal{F}$, then we can just let $g(x) = k$ and $X = \Omega$ (of course, the BCT will also be valid for any $X \in \mathcal{F}$, just as the DCT). We know that k is integrable since it is just a constant and can be represented by a simple function. Anyway, we now see that the theorem must follow by the DCT. □

Another important named theorem that follows from the DCT is something called *Scheffe's theorem*. It tells us something nice about collections of nonnegative measurable functions. The proof follows the one presented on page 64 in Athreya's book [Ath].

Theorem 3.4.7. *Let (f_n) be a collection of nonnegative measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$. Also, suppose that this sequence converges pointwise to a function f almost everywhere in Ω. Also, suppose that $\int_\Omega f_n \, d\mu \to \int_\Omega f \, d\mu$ and $f \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Then,*

$$\lim_{n\to\infty} \int_\Omega |f_n - f| \, d\mu = 0.$$

Proof. Let us construct a new sequence of functions (g_n). We will say that for each $n \in \mathbb{N}$, $g_n = f - f_n$. Then, since $f_n \to f$ almost everywhere, we see that $g_n = g_n^+ - g_n^-$ must go to zero almost everywhere, and so each function g^+ and g^- must go to zero almost everywhere. Again, recall that the superscripts on these functions simply denote the positive and negative parts of the function g. This kind of language is analogous to the language we employed when defining the integral of an arbitrary measurable function (see Def. 3.3.5). Anyway, we see that we also have $0 \leq g_n^+ \leq f$, because the f_n's are nonnegative functions and we defined $g_n = g_n^+ - g_n^- = f - f_n$ which means that $0 \leq g_n^+ \leq g_n \leq f$, since g_n^+ is nonnegative by definition. Anyway, we also have $\int_\Omega f \, d\mu < \infty$ by our hypothesis. So, we have a sequence of functions g_n^+ converging pointwise to 0, bounded above by the function f. Therefore, we can apply the DCT to find that

$$\int_\Omega g_n^+ \, d\mu \to 0.$$

However, we also hypothesized that $\int_\Omega f_n \, d\mu \to \int_\Omega f \, d\mu$. Therefore, $\int_\Omega g_n \, d\mu = \int_\Omega f_n \, d\mu - \int_\Omega f \, d\mu \to 0$. It follows from the definition of g_n^- that $\int_\Omega g_n^- \, d\mu = \int_\Omega g_n^+ \, d\mu - \int_\Omega g_n \, d\mu \to 0$. Now that we have shown the convergence of the integrals of our positive and negative parts of g_n, we conclude that

$$\int_\Omega |g_n| \, d\mu = \int_\Omega |f_n - f| \, d\mu = \int_\Omega g_n^+ \, d\mu + \int_\Omega g_n^- \, d\mu \to 0 \text{ as } n \to \infty.$$

We are now done with the proof! □

We conclude this section by mentioning that the named theorems we have shown above are very powerful. They tell us about the behavior of integrals we might have no idea how to compute, given that we know something about the integrand. For example, the DCT tells us that if we can construct a sequence of functions that converge to some limiting function and that are bounded by an integrable function, then we will be able to say something about the integrals of these functions and the integral of the limiting function. Indeed, we can say that the limit of the integrals of the sequence of functions converges to the integral of the limiting function. This may seem like a

nonsense sentence, but what we are really saying is that we are able to exchange the limit and integral operations. Again, this is something we often do (especially naive physicists), without ever thinking about the powerful mathematical formalism that we are invoking. We illustrated in this section the difficulty and depth of the mathematical rigor behind such seemingly innocuous exchanges. Also, we have shown that these convergence theorems are very powerful tools because we sometimes only have limited information about a particular function and are forced to approximate it with sequences of other functions.

3.5 Convergence

Now, the reader may have noticed that we have discussed many different kinds of convergence in the previous sections of this paper. Indeed, it is often very confusing to think about what kind of particular convergence one is talking about. However, it is very important that such things are well-defined. Convergence will be a fundamental issue in the sections that follow. Thus, in order to eliminate the confusion, the objective of this section is to put in one place all of the different kinds of "convergences". Specifically, we want to list all of the different ways a sequence of measurable functions (f_n) converges to a limiting function f. So, in the definitions that follow, we will always assume that (f_n) is a sequence of measurable functions on $(\Omega, \mathcal{F}, \mu)$. Let us begin.

Definition 3.5.1. We say that (f_n) converges to f pointwise if given any $x \in \Omega$, we have the following:

$$\lim_{n\to\infty} f_n(x) = f(x) \text{ for all } x \in \Omega.$$

Indeed, pointwise convergence simply means that the sequence of real numbers $(f_n(x))$ (given by the *value* of the functions f_n at each point $x \in \Omega$) converges to the real number $f(x)$.

Definition 3.5.2. (f_n) converges to f *almost everywhere* in Ω if there exists a set $A \in \mathcal{F}$ such that $\mu(A) = 0$ and

$$\lim_{n\to\infty} f_n(x) = f(x) \text{ for all } x \in A^{\mathcal{C}}.$$

In other words, convergence almost everywhere is simply pointwise convergence on all points in $\Omega \setminus A$, where A is some subset of Ω with measure 0.

Definition 3.5.3. We say that (f_n) converges to f in $\mathcal{L}(\Omega, \mathcal{F}, \mu)$ if $\int_\Omega |f_n|\, d\mu < \infty$ for all $n \geq 1$, $\int_\Omega |f|\, d\mu < \infty$, and

$$\lim_{n\to\infty} \int_\Omega |f_n - f|\, d\mu = 0.$$

We will denote such a convergence by $f_n \to f$ in $\mathcal{L}(\Omega, \mathcal{F}, \mu)$.

4 Probability

4.1 Kolmogorov's Probability

Of course, as we mentioned previously, the mathematical language we have been using has direct analogues to probabilistic language. Indeed, the notions of measure spaces, measurable functions, and integrals have their counterparts in probability theory. The analogy between these measure theoretic concepts and probability theory was first articulated by Kolmogorov in 1956 [Ath]. The title of this subsection reflects this historical fact. So, to begin, let us redefine our notion of a measurable space a little bit. So, we call that set Ω a *sample space*. This is a set of all possible outcomes of some sort. Therefore, we call each element $x \in \Omega$ a *sample point*. Now, it may be the case that many sample point all have some characteristic property. For example, when we throw a six-sided die, there are many orientations of the die that yield a six on top. Indeed, we can throw the die so that the six mark is facing us, or is side-ways, or is upside down, etc. However, we want to be able to characterize a "six on top" as an *event*. Thus, an event is a subset of our sample space Ω. Then, we can naturally think about our σ-algebra $\mathcal{F}$ as the set of all events. So, what is our measure μ, then? It is just the probability of an event! Indeed, we are able to say that for any event $X \in \mathcal{F}$, $\mu(X)$ represents the probability of this event occurring. The only difference between this probabilistic μ and the measure μ we have been talking about is that we *require* that $\mu(\Omega) = 1$.

Now, before we continue, we have to make sure that our properties of a σ-algebra make sense for the space of all possible events. So, let us go through the different properties of σ-algebras that we had.

i. $\Omega \in \mathcal{F}$: This condition makes sense because we need to be able to say that any sample point in our sample space is something that occurs. It does not necessarily *ever have to occur*, but it must be such that we can measure its probability of occurring. Indeed, this is how we usually *define* our sample space. For example, when we list all the possible ways that we can throw a six-sided die and put them in Ω, we do not include something silly like the set of all trees in Cleveland. We cannot measure the probability of a tree! Thus, we require that we are able to measure the probability of our entire sample space. Further, since our sample space lists all

the possible events that occur, we also know that $\mu(\Omega) = 1$.

ii. $A \in \mathcal{F}$ implies $A^{\mathcal{C}} \in \mathcal{F}$. This assumption is basically saying that if we can measure the probability of an event *occuring*, then we can measure the probability of the event *not occuring*. Again, this is a very reasonable assumption.

iii. $A, B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$. This assumption means that if we can measure the probability of event A occurring, and if we can also measure the probability of event B occurring, then we can certainly measure the probability of *either* of the events occurring. Of course, probability theory must include such an assumption if we want to look at multiple events, such as the probability of rolling either a two or a three on a six-sided die.

iv. $A_n \in \mathcal{F}$ for $n \geq 1$ implies $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ and $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$. This assumption allows us to not limit ourselves to finite numbers of outcomes. Indeed, sometimes we want to be able to find the probability that any (union) of a countable number of events occurs, or the probability that *all* (intersection) the events occur. A very simple example of a probabilistic experiment that requires such a countable infinity is the tossing of a two-sided coin until the first head comes up ([Ath] pg. 190). Indeed, although the chances are very small, we can toss the coin an arbitrarily large number of times before we get the first heads. Thus, the nature of probabilistic events requires us to make this assumption.

So, now that we have gone through all the relevant properties of σ-algebras, we see that it really does make sense that each of our events X are elements of a σ-algebra $\mathcal{F}$ associated with Ω. There is still the matter, however, of the properties associated with the measure μ. We have to make sure that these properties make sense in the context of probability. Let us now list these properties and discuss them.

i. $\mu(A) \in [0, 1]$ for all $A \in \mathcal{F}$. Now, this assumption is basically saying that we do not allow for negative probabilities. Also, we cannot have a probability greater than 1 since we already know that $\mu(\Omega) = 1$, and $A \subseteq \Omega$.

ii. $\mu(\emptyset) = 0$. This basically says that if our event has no associated sample point in it, that is, the subset of Ω is empty, then the probability of that event occurring is zero. This is the same as thinking about the probability of a seven occurring on a six-sided die. We know that since a seven is not even printed on the die, then there are *no sample points in* Ω that correspond to such an event. Therefore, this subset in $\mathcal{F}$ is an empty set and we expect that $\mu(\emptyset) = 0$.

iii. For any pairwise disjoint collection of sets $A_1, A_2, \ldots \in \mathcal{F}$ with $\bigcup_{n\geq 1} A_n \in \mathcal{F}$, $\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$. Notice that disjoint events in probability are *mutually exclusive*. Indeed, if for two events A and B, $A \cap B = \emptyset$, then there are no possible ways of the two events both occurring. Therefore, it makes sense that the probability of *either* of these two mutually exclusive events occurring is just the sum of the individual probabilities: $\mu(A \cup B) = \mu(A) + \mu(B)$. We can also think about this in terms of our die example. Rolling a two and rolling a one are two mutually exclusive events. Thus, the probability of rolling either is just $1/6 + 1/6 = 1/3$, just as we expect. So, this property is allowing us to make such statements about probabilities. It also allows us to not limit ourselves to finite numbers of outcomes. Indeed, we see that this mutual exclusivity condition is being applied to a countable infinite number of sets.

So, we have now gone through the definitions of $\mathcal{F}$ and μ and found that each property they satisfy has a direct correlation to a natural property of either an event or a probability. Thus, we can now move on to define some other notions in probability theory that have direct connections to other measure theoretic concepts we described.

The next interesting definition we will have is that of a *random variable*.

Definition 4.1.1. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $X : \Omega \to \mathbb{R}$ be a measurable function. Then, X is called a *random variable* on $(\Omega, \mathcal{F}, \mu)$.

Thus, measurable functions from sample spaces to the real numbers are random variables! This should make intuitive sense. For example, the outcome of a six-sided die: 1, 2, 3, 4, 5, 6 is a random number. But, what does this mean? This means that for every possible throw of our die (which may include all of the different physical determinants of a die roll, such as wind speed, friction between the die and table, etc.), that is for every sample point x in our sample space Ω, there is an associated result: $1, 2, 3, 4, 5, 6$ in the real numbers. Therefore, the outcome of the die roll is really a function that maps all the possible sample points to a particular numeric result. The condition that such a function is measurable makes sense because we know that outcomes behave nicely with respect to the subsets of Ω. What we mean by this is that an outcome of a six-sided die, say, corresponds to the event (subset of Ω) that a six was rolled. Thus, random variables *must be associated with subsets of Ω whose probability we can measure*. To drive this point home, let us think about our condition for measurable functions f. We have that for each $\alpha \in \mathbb{R}$, $\{x \mid f(x) \leq a\} \in \mathcal{F}$. This is simply saying that if we take a particular random number, a, say, and if we look at the sample points associated with that number, then the set of all such sample points that map to a (or anything less than a, for

that matter) has to be a subset of Ω whose probability we can measure. In other words, we expect that for any given range of our function, the associated domain is in $\mathcal{F}$. So, now that we have our definition of a random number, we can do all sorts of fun things because random variables are just measurable functions! Indeed, given that we already went through all that trouble defining integrals of measurable functions, we can now apply the same thing to random variables.

4.2 Random Variables

From now on, we will refer to random variables as X. This is a very common convention. However, it is important to keep in mind that X is really a *function*. There is nothing particularly variable about it! Indeed, the reason we call it variable is because in a probabilistic experiment, we usually have no idea what particular sample point in Ω maps to a particular event. Indeed, all we know is the value of X. Anyway, the first thing to look at is the expected value of a random variable. This is the value that we get from the variable *on average*.

Definition 4.2.1. Let X be a random variable on $(\Omega, \mathcal{F}, \mu)$. The *expected value* of X, denoted by $E(X)$, is defined as

$$E(X) = \int_\Omega X \, d\mu,$$

provided the integral is well defined.

Let us give a very easy example in terms of our six-sided die. Let X be the random variable that describes the numeric outcome of a die roll. This means that X can only take on the values 1 through 6. This is great because this means that X is not *just* a measurable function, but it is a *simple function*! So, to compute our integral, let's partition our sample space Ω into 6 partitions $A_1, A_2, \ldots A_6$ that correspond to the sample points that yield a roll of one, two, etc. Notice that since two of these events can't possibly happen together, the A_i's are mutually disjoint and partition Ω. Therefore, our random variable X can be written as the sum

$$X = \sum_{i=1}^{6} i \, I_{A_i}.$$

We now have to make an assumption about our sets A_i. We will suppose that our die is fair. This means that there are equally many possible die rolls that result in a one, or a two, or a three, etc. In other words, we are saying that the die isn't weighted and that the conditions are such that

no side of the die has more possible rolls associated with it. Anyway, if this is the case, then we expect the sizes (probabilities) of the sets of sample points associated with each outcome to be the same. Mathematically, then, $\mu(A_1) = \mu(A_2) \ldots = \mu(A_6)$. However, since these A_i's partition Ω, $\mu(\bigcup_{i=1}^{6} A_i) = \mu(\Omega) = 1 = \sum_{i=1}^{6} \mu(A_i)$. Consequently, $\mu(A_i) = 1/6$ for all $i = 1, 2, \ldots, 6$. Now that we have established the probabilities of these events, we can use our definition of the integral of a simple function to compute the expected value of the die roll:

$$E(X) = \int_{\Omega} X \, d\mu = \sum_{i=1}^{6} i\mu(A_i) = \frac{1}{6} \sum_{i=1}^{6} i = \frac{21}{6} = 7/2$$

We can think about the number $E(X)$ in another way. Suppose we throw our six-sided die n times. We can consider each toss X_i as being a random variable, i.e. a measurable function from the sample space Ω to the real numbers. Then, we construct a sequence of functions (S_n) as follows. Given any $n \in \mathbb{N}$, $S_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_n(x)$ for every $x \in \Omega$. The sequence (S_n) must be a sequence of random variables because we stated in Sec. 2.3 that the finite sum of measurable functions is a measurable function. Here S_n is the random variable that corresponds to the average of n die rolls, or the number we get by writing down n die roll results and dividing their sum by the total number of rolls. We may now say that our average of the rolls will converge in probability to $E(X)$. This means that for every $\epsilon > 0$, $\mu(\{x \in \Omega : |S_n(x) - E(X)| < \epsilon\}) \to 1$ as $n \to \infty$ (where $E(X) = 7/2$). In other words, as we toss the die more and more often, the chance that the average of the rolls is close to $7/2$ becomes almost certain (approaches probability 1). Such a convergence is called a *weak law of large numbers*. This is a fundamental result in probability theory. However, it is tangential to our discussion and so we will not prove it. We refer the interested reader to page 238 in [Ath] for an outline of the argument.

Some other important definitions in probability theory include those of a *probability distribution function* and *cumulative distribution function*, or cdf, of a random variable X. We shall define these now.

Definition 4.2.2. Let X be a random variable on $(\Omega, \mathcal{F}, \mu)$. Let

$$F_X(\alpha) \equiv \mu(\{x \in \Omega \mid X(x) \leq \alpha\}), \text{ where } \alpha \in \mathbb{R}.$$

Then we call the function F_X a *cumulative distribution function* (cdf) of X.

Definition 4.2.3. Let X be a random variable on $(\Omega, \mathcal{F}, \mu)$. Let

$$P_X(A) \equiv \mu(X^{-1}(A)) \text{ for all } A \in \mathcal{B}(\mathbb{R}).$$

Then we call the function P_X a *probability distribution function* (pdf) of X.

Notice that these definitions simply describe probabilities of particular random numbers occuring. For the cdf, we are looking at the probability that our random number X is less than or equal to some number $\alpha \in \mathbb{R}$. For the pdf, notice that we are simply mapping nicely behaved subset of the real numbers (the Borel sets) that our function X may take on to the subset of Ω that corresponds to this Borel set ($X^{-1}(A)$). Then, the pdf is just the probability that our random variable X is one of the values in this Borel set. It makes sense to define a pdf because we often do not have a discrete random variable. For instance, we can think about throwing a baseball into the air and measuring the maximum height it attains with respect to the ground. This will naturally be a random variable because we cannot possibly control all of the conditions that determine the height of the ball. Also, the actual value of the height will be some positive real number. Therefore, it is not useful to talk about the probability of a particular number occurring, such as 2.233 meters. There are uncountably many real numbers, and the probability (measure) of a particular value occurring is zero. However, it is possible to assign a non-zero probability to a range of heights. Indeed, we can consider the probability that the ball reaches a height between 2.3 and 2.4 meters. This is analogous to finding the pdf of the Borel set $A = [2.3, 2.4]$.

One of the ways probability differs from measure theory is in the other ways we can describe events. Indeed, we are not limited to just looking at events and their unions, complements, and intersections. An important concept in probability is the idea of *independent events*. We will define them using the language in Athreya's book on page 219 [Ath].

Definition 4.2.4. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $\{A_\alpha \,|\, \alpha \in \Lambda\} \subset \mathcal{F}$ be a collection of events (Λ denotes some arbitrary indexing set). We call these events *independent* with respect to μ if for every *finite* subcollection $\{A_{\alpha_1}, A_{\alpha_2}, \ldots, A_{\alpha_n}\}$,

$$\mu\left(\bigcap_{j=1}^{n} A_{\alpha_j}\right) = \prod_{j=1}^{n} \mu(A_{\alpha_j}).$$

Also, we may have the case that we only have a finite collection of events $\{A_1, A_2, \ldots, A_n\}$. Then, if $\mu(A_i \cap A_j) = \mu(A_i)\mu(A_j)$ for each i and j such that $i \neq j$, we say that our sets in the collection are *pairwise independent*.

In order to illustrate these definitions, suppose that we have two events $A, B \in \mathcal{F}$. Then, suppose that we know the event B occurred. Now we want to know what is the probability that event A

occurred *given* that event B occurred. Clearly this has to be the ratio of the probability that *both* the events occur to the probability that event B occurs. Mathematically, we write this fraction as $\mu(A \cap B)/\mu(B)$. Now, if the two events are independent, then we know that this ratio (also called the conditional probability of A given B) is just equal to $\mu(A)$. This means that the probability of A occurring given B occurred is the same as the probability of A occurring *independent of our knowledge of* B. So, we now have our conceptual definition of independence. We say that events are independent when the knowledge about the outcome of one event tells us nothing about the outcome of another event. Finally, we may extend the concept of independence to random variables. In this case, we say that two random variables $X : \Omega \to \bar{\mathbb{R}}$ and $Y : \Omega \to \bar{\mathbb{R}}$ are independent if for any real numbers $a, b \in \bar{\mathbb{R}}$, the events $\{x \in \Omega : X(x) \leq a\}$ and $\{x \in \Omega : Y(x) \leq b\}$ are independent according to Def. 4.2.4 above (pg. 221 in [Ath]).

Before we move on to ergodic theory, it is important to emphasize the idea of convergence in probability theory. We already talked about this in the die roll example described above. However, in order to facilitate later proofs, we present the following definition.

Definition 4.2.5. Suppose we have a sequence (X_n) of random variables on a probability space $(\Omega, \mathcal{F}, \mu)$. Then, we say that this sequence converges with probability 1 to a random variable X if there exists a set $A \in \mathcal{F}$ such that

$$\mu(A) = 1 \text{ and for all } x \in A, \lim_{n\to\infty} X_n(x) = X(x).$$

In other words, we say that $X_n \to X$ with probability 1, or (X_n) converges to X pointwise almost everywhere in A.

5 The Ergodic Theorem

5.1 Transformations

We finally move on to ergodic theory. In very broad terms, ergodic theory is the study of transformations of measure spaces. In all of our previous discussions, we have assumed that our spaces $(\Omega, \mathcal{F}, \mu)$ remain static. Indeed, if we look at the statements of all the theorems we have proven, we have always stated something like this at the beginning: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Now we want to be able to take our space Ω, and map it into itself via some function $T : \Omega \to \Omega$. Of

course, if we do this arbitrarily, we can really ruin all of our underlying measure structure. Indeed, it might be the case that the transformation T will map subsets of Ω that used to be in our σ-algebra $\mathcal{F}$ to subsets that are not in the σ-algebra! This would be disastrous because we would lose our ability to measure these subsets with our measure μ. Thus, ergodic theory concerns itself with those transformations that preserve the measure theoretic structure of the space. Indeed, we have to make sure that the transformations are *measurable*. Let us now define this term.

Definition 5.1.1. Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space. Then, we call the function $T : \Omega \to \Omega$ a *measurable transformation* if

$$T^{-1}(A) \in \mathcal{F} \text{ for all } A \in \mathcal{F}.$$

However, we are not quite finished. Just because we know that the transformation preserve the measurability of our subsets of Ω, it doesn't mean that the actual sizes of the subsets remain the same as we apply the transformation T on them. Therefore, ergodic theory requires us to look at *measure-preserving* transformations. These transformations not only conserve the structure of $\mathcal{F}$, but also make sure that the structure of our measure μ is preserved. Specifically, if we look at any particular subset $A \in \mathcal{F}$, then all of the points in Ω that are mapped to this subset by T, i.e. $T^{-1}(A)$, must have the same size as A. In other words, we do not stretch or shrink our measure on Ω by applying the transformation. This is illustrated in Fig. 7. To make this notion of measure-preserving transformations a little more formal, we have the following definition. From Athreya's book on page 40:

Definition 5.1.2. First, let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $T : \Omega \to \Omega$ be a measurable transformation. Then, T is called *measure preserving* on $(\Omega, \mathcal{F}, \mu)$ if for all $A \in \mathcal{F}$, $\mu(T^{-1}(A)) = \mu(A)$.

To test out these issues, let us do a little exercise suggested by Athreya on page 272 [Ath]. Suppose that we have $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$, and we let μ be the Lebesgue measure. Then, we define our transformation in the following way

$$T(x) = \begin{cases} 2x & \text{if } 0 \leq x < \frac{1}{2} \\ 2x - 1 & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{if } x = 1 \end{cases}$$

An illustration of this transformation is given in Fig. 8. We see in the figure that the transformation T maps the interval $[0, 1]$ into itself. In doing so, T "stretches" the two halves of the interval into

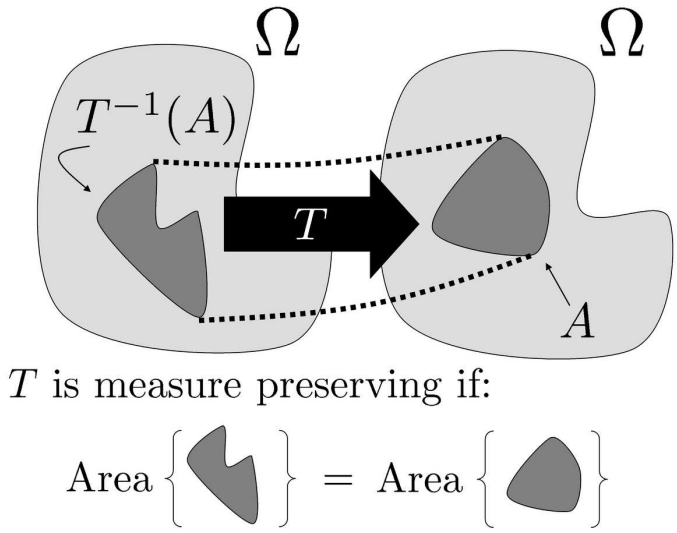

Figure 7: *This illustrates what we mean, conceptually, by a measure-preserving transformation T on a measure space $(\Omega, \mathcal{F}, \mu)$.*

"strips". Although only a single point is transformed in the figure, the reader may iterate other points in the $[0, 1]$ interval by using a straightedge and performing the graphical analysis illustrated for point z in Fig. 8. The collection of points $\{z, T(z), T^2(z), \ldots\}$ for some $z \in [0, 1]$ is called the *orbit* of z under T.

We want to know if T is a measure-preserving transformation on $[0, 1]$. Recall from our discussion of the Borel σ-algebra, that $\mathcal{B}$ is the σ-algebra that is generated by open subsets of $\mathbb{R}$. In this case, we are only looking at the Borel σ-algebra generated by open subsets of $[0, 1]$. So, to show that T is measure-preserving, we will prove that T preserves the measure of all open subintervals of $[0, 1]$. This is sufficient because we know from our discussion in Sec. 2.2 that the Lebesgue measure of a subset of the real numbers does not change if we remove or add any countable set of real numbers to the subset. Therefore, showing that the measures of open intervals in $[0, 1]$ are preserved implies that the measures of all other intervals in $[0, 1]$ are preserved, as well. Further, by the additivity property of the measure, we know that all other measurable subsets of $[0, 1]$ will have a preserved measure under this transformation T. To summarize, if the measures of all open subintervals of $[0, 1]$ are preserved, then so are the measures of all Borel subsets of $[0, 1]$. For brevity, we have omitted a formal proof of this fact and simply discussed the reasons why this is the case.

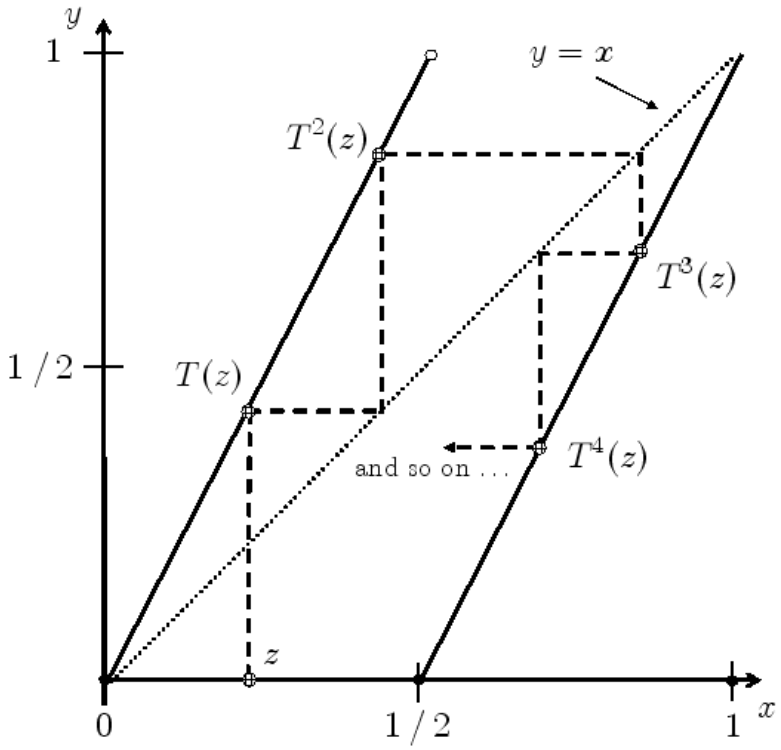

Figure 8: *This is an illustration of the transformation T defined in the exercise on page 272 in [Ath]. The point z is an example of a "starting position" for the transformation T. This figure shows the first few points in the orbit of z under T (checkered points). The dotted $y = x$ line provides a visual guide for performing these iterations.*

Hence, we will only consider open intervals (a, b) where $0 < a < b < 1$. We already know that $\mu((a, b)) = b - a$ from our discussions of the Lebesgue measure. So, suppose that our transformation T mapped some subset of the real numbers to an interval (a, b). What is $T^{-1}((a, b))$? Well, it has to be the set of all $x \in [0, 1]$ such that $a < T(x) < b$. Now, notice that $T(x) = 2x$ for all $0 \leq x < 1/2$. Therefore, the interval $[0, 1/2)$ gets mapped to $[0, 1)$. This is the first dark diagonal line on the left-hand side of Fig. 8. Hence, since $(a, b) \subset [0, 1)$, then we know that $(a/2, b/2) \subset T^{-1}((a, b))$. The other possibility is that $1/2 \leq x < 1$. In this case, we know that $T(x) = 2x - 1$. Therefore, our interval $[1/2, 1)$ get's mapped to $[0, 1)$, also. This is the second dark diagonal line from the left in Fig. 8. We see again that since $(a, b) \subset [0, 1)$, $(a/2 + 1/2, b/2 + 1/2) \subset T^{-1}((a, b))$. Now, the other possibilities for x are just the endpoints of the interval $[0, 1]$. However, $T(1) = 0$ and $T(0) = 0$. Consequently,

these end points cannot be in $T^{-1}((a,b))$ since $0 < a < b < 1$. Anyway, we have now found the regions in $[0,1]$ that map to (a,b) under T. But, consider that the intervals $(a/2 + 1/2, b/2 + 1/2)$ and $(a/2, b/2)$ are mutually disjoint since $0 < a < b < 1$. Therefore,

$$\begin{aligned}\mu(T^{-1}((a,b))) &= \mu\left(\left(\frac{a}{2},\frac{b}{2}\right) \cup \left(\frac{a}{2}+\frac{1}{2},\frac{b}{2}+\frac{1}{2}\right)\right) \\ &= \mu\left(\left(\frac{a}{2},\frac{b}{2}\right)\right) + \mu\left(\left(\frac{a}{2}+1,\frac{b}{2}+1\right)\right) = \frac{b}{2} - \frac{a}{2} + \frac{b}{2} + \frac{1}{2} - \frac{a}{2} - \frac{1}{2} = b - a!\end{aligned}$$

We have now shown that $\mu(T^{-1}((a,b))) = \mu((a,b))$. Thus, T must be a measure-preserving transformation. So, in this example we have seen how to figure out if a transformation is measure preserving. Indeed, we have to consider very carefully what subsets of our set Ω map to the particular $A \in \mathcal{F}$ we are looking at. As we have seen in this example, it can often be that two disjoint subsets map to the particular subset. The measure T can split and join together subsets of Ω as it pleases, *as long as the total area of these subsets remains the same* as we transform our set Ω.

We now need one more condition on the transformation T that allows us to talk about *ergodicity*. This is the notion of an *ergodic* transformation. We will first define it and then talk about it.

Definition 5.1.3. Let T be a measure preserving transformation on a probability space $(\Omega, \mathcal{F}, \mu)$. A set $A \in \mathcal{F}$ is T-*invariant* if $A = T^{-1}(A)$. Now, if A is T-invariant implies that $\mu(A) = 0$ or $\mu(A) = 1$, then we say that T is *ergodic* with respect to μ.

At this point, we will not be able to say much conceptually about these ergodic transformations. Indeed, right now, we can only leave this concept at the level of the definition. We will require the powerful Ergodic Theorem to make some conceptual statements about ergodic transformations. However, before we leave this definition alone, we can give an easy example of a transformation that is measure preserving, but not ergodic! This example is found in Athreya's book on page 273. Anyway, suppose that our sample space $\Omega = \{x_1, x_2\}$, such that $x_1 \neq x_2$. Then, let us define our transformation $T : \Omega \to \Omega$ such that $T(x_1) = x_2$ and $T(x_2) = x_1$. Now, suppose that our measure μ is just $\mu(\{x_1\}) = \mu(\{x_2\}) = 1/2$. So, it is almost trivial to show that T is measure preserving. Indeed, this can be easily checked because we have so few possible subsets of Ω. In fact, we can specify the entire powerset of Ω: $\mathcal{P}(\Omega) = \{\emptyset, \{x_1\}, \{x_2\}, \Omega\}$. We see that $\mu(T^{-1}(\{x_1\})) = \mu(T^{-1}(\{x_2\})) = \mu(\{x_1\}) = \mu(\{x_2\}) = 1/2$. Further, $\mu(T^{-1}(\Omega)) = \mu(\Omega) = 1$ and $\mu(T^{-1}(\emptyset)) = \mu(\emptyset) = 0$. We can also easily show that T is ergodic. Notice that the only subsets of Ω that map to themselves are the empty set and Ω, itself. In the other cases, we have that $T(\{x_1\}) = \{x_2\}$ and vice-versa.

Anyway, since $\mu(\emptyset) = 0$ and $\mu(\Omega) = 1$, then we know that T is ergodic. What about T^2, i.e., the transformation $T^2(x) = T(T(x))$ for all $x \in \Omega$? Notice that T^2 must be the identity transformation because $T(T(x_1)) = T(x_2) = x_1$ and $T(T(x_2)) = T(x_1) = x_2$. However, this means that every subset of Ω is mapped to itself! Consequently, although $T^2(\{x_1\}) = \{x_1\}$, we have that $\mu(\{x_1\}) = 1/2$, which is not equal to either 0 or 1. Thus, T^2 is measure-preserving, but not ergodic!

Now, these last considerations have been very simplistic. However, we can actually use the previous example to motivate the Ergodic Theorem! Specifically, notice that if we iterate our transformation T, we manage to cover both of our sample points $x_1, x_2 \in \Omega$. In other words, what we mean is if we start with either x_1 or x_2, then $T(x_1) = x_2$, and $T(T(x_1)) = T^2(x_1) = x_1$, and $T^3(x_1) = x_2$, and so forth. Thus, we see that T *samples the entire sample space* Ω under iteration. What about the transformation T^2? This function does *not* sample our entire sample space under iteration. Explicitly, $(T^2)^n(x_1) = x_1$ and $(T^2)^n(x_2) = x_2$, for all $n \in \mathbb{N}$. Thus, as we iterate T^2, we do not get to all of the elements in Ω. Now, it is *not coincidental* that T has this sampling property and is ergodic, while T^2 does not sample our space very well and is not ergodic. As we shall see in the next section, ergodicity and sampling are intimately related. Indeed, it will be shown (albeit in a more formal language) that ergodic maps $T : \Omega \to \Omega$ have the property that the orbit of any "initial condition" $x_0 \in \Omega$ under T will sample the space Ω arbitrarily well. We will clarify what we mean by this in the next section.

5.2 Birkhoff's Ergodic Theorem

We will begin with a statement of Birkhoff's Ergodic Theorem. Originally, this theorem, described by Birkhoff in 1931, dealt with differential equations [Bi]. The theorem characterized the behavior of trajectories described by the solutions to these differential equations. In our case, the trajectories will be described by iterations of an ergodic transformation T. Like Birkhoff, we will be interested to see what happens to random variables under these trajectories, or iterated transformations. Anyway, the version of our theorem, as stated in Athreya's book on page 274, goes as follows:

Theorem 5.2.1. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, $T : \Omega \to \Omega$ be an ergodic transformation and $X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Consider the sequence of functions (f_n) (where $f_n : \Omega \to \bar{\mathbb{R}}$ for each $n \in \mathbb{N}$) given by*

$$f_n(x) = \frac{1}{n} \sum_{j=0}^{n-1} X(T^j(x))$$

for all $n \in \mathbb{N}$*. Then, for each* $x \in \Omega$*, the sequence of real numbers* $(f_n(x))$ *converges to the expectation of the random variable* X*, i.e.* $E(X)$*. In other words, for our sequence of functions* (f_n)*,*

$$f_n(x) = \frac{1}{n}\sum_{j=0}^{n-1} X(T^j(x)) \to E(X) \equiv \int_\Omega X \, d\mu$$

with probability 1 (see Def. 4.2.5) and in $\mathcal{L}$ *(see Def. 3.5.3) as* $n \to \infty$*. As a final clarification, here we have a pointwise convergence of* (f_n) *almost everywhere to a constant function that has the value* $E(X)$ *everywhere in* Ω*.*

Now, let us discuss the elements of this theorem a little bit. First of all, we see that we are dealing with some random variable X. Recall that a random variable is just a real-valued measurable function on our outcome space Ω. However, we specify in the theorem that $X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Recall from Def. 3.3.6, that this implies that the integral of $|X|$ over Ω is finite. Indeed, this is what allows us to take the integral of X to get an expected value. This brings us to one of the quantities we are looking at in this theorem. Notice that on the right-hand-side of the equality in the theorem we just have $E(X)$. This is just the average value of X over every point in Ω. In physics, this kind of average is called a *phase space* average. Indeed, a phase space in physics is just the set Ω of all the possible states of a particular system. Then, if we examine the value of some variable X at each point in the phase space, the average value that this variable X takes on is just the average $E(X)$.

We also have in the statement of Birkhoff's ergodic theorem an interesting term denoted by

$$\frac{1}{n}\sum_{j=0}^{n-1} X(T^j(x)).$$

Now, notice that for each $n \in \mathbb{N}$, the quantity above is a function from Ω to the real numbers. This is the case because we know that $X : \Omega \to \mathbb{R}$ is a random variable. Moreover, each f_n must be a random variable as well, since it is a finite sum of random variables (measurable functions). Now, look at the argument of our function X: $T^j(x)$. We know that our transformation T maps points in Ω to other points in Ω. Thus, $X(T^j(x))$ represents the value of X at a point in the orbit of x under T. Indeed, we are looking at the behavior of the random variable X as we allow our system to be transformed under the iteration of the ergodic T transformation. So, now that we have worked out $X(T^j(x))$, consider the sum. We are basically saying, okay, take a point $x \in \Omega$, and compute the value of X at x, $T(x)$, and so forth. Then, add up the values of X at these points, i. e. perform the sum: $\sum_{j=0}^{n-1} X(T^j(x))$. Finally, divide this sum by the total number of iterations performed. This

is what gives us the factor of $1/n$. Thus, what we are doing here is computing a *time average* of X, where "time" is represented by the number of times we iterate the map T. Now that we have described the two important terms of the theorem, let us really see what this Ergodic Theorem is saying.

In broad terms, the theorem is saying that the *time average* of X converges to the *phase space* average as $n \to \infty$. So, basically, if we iterate our transformation T at some $x \in \Omega$, then the average we compute for the variable X over the points in the orbit of x under T will approach the expected value of the random variable! Finally, it is important to note that the function $\frac{1}{n}\sum_{j=0}^{n-1} X(T^j(x))$ converges to $E(X)$ at all points $x \in \Omega$. Therefore, the nice property of our time average *is independent* of our starting position $x \in \Omega$. So, now that we've tried to state the idea behind the theorem in many ways, consider Fig. 9 that tries to illustrate this concept.

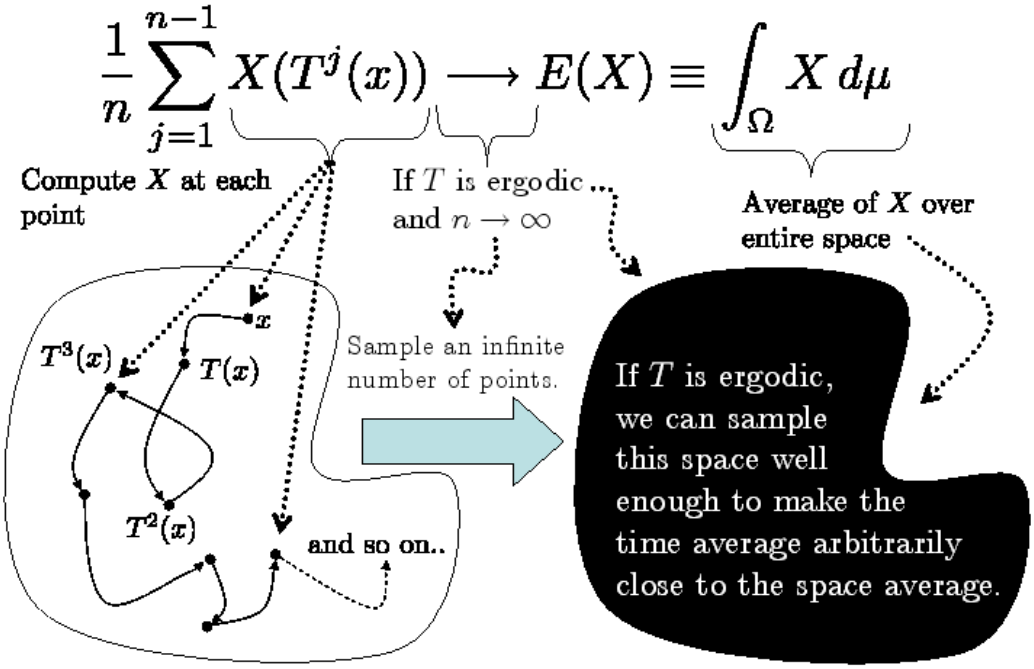

Figure 9: *This is an illustration of what the Ergodic Theorem means, conceptually. The top half of the figure just lists the relevant mathematical terms in the theorem. The bottom half are a graphical representation of the procedure for computing the terms in the top half. The dotted arrows connect the mathematical terms to their corresponding concepts. Note that, in this figure, the blobs represent the space* Ω.

After this explanation, we will proceed down the road to actually proving this theorem. The first step will be to show an important lemma. The statement and proof of the lemma is given on page 274 in Athreya's book. We will follow the proof given on that page.

Lemma 5.2.2. *(Maximal ergodic inequality) Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$. Next, suppose that we have an integral random variable $X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Let us now construct a sequence of functions $S_0, S_1, \ldots$ in the following way. Suppose that $S_0(x) = 0$ for all $x \in \Omega$. Next, suppose that $S_n(x) = \sum_{j=0}^{n-1} X(T^j(x))$ for all $n \in \mathbb{N}$. Finally, consider another sequence of functions $M_0, M_1, \ldots$ where $M_n(x) = \max\{S_j(x) : 0 \leq j \leq n\}$. Then, we have that for each $n \in \mathbb{N}$,*

$$E(X(x) I_{\{y \in \Omega \,:\, M_n(y) > 0\}}(x)) \geq 0.$$

Proof. Let $n \in \mathbb{N}$. If $M_n(x) \leq 0$ for all $x \in \Omega$, then from the definition of a characteristic function, $I_{\{y \in \Omega : M_n(y) > 0\}}(x) = 0$ for all $x \in \Omega$, and the theorem trivially follows. Thus, suppose that this is not the case and consider some point $x \in \Omega$ such that $M_n(x) > 0$. Anyway, since we defined our functions M_n such that they are the maximum value of the functions S_j for all $0 \leq j \leq n$ (at any particular $x \in \Omega$), then we know that $M_n(x) \geq S_j(x)$ for all $1 \leq j \leq n$. Further, since this inequality is true for any $x \in \Omega$, and $T(x) \in \Omega$, $M_n(T(x)) \geq S_j(T(x))$, as well. Therefore, adding the value of the random variable at $x \in \Omega$ to both sides of the inequality yields

$$X(x) + M_n(T(x)) \geq X(x) + S_j(T(x)) = S_{j+1}(x),$$

for all $0 \leq j \leq n$, where the last equality follows from the definition of the S_j's. We now have that $X(x) \geq S_{j+1}(x) - M_n(T(x))$ for all $0 \leq j \leq n$. We can now reindex our counter j and conclude that $X(x) \geq S_i - M_n(T(x))$ for all $1 \leq i \leq n$. Therefore, by the definition of the maximum, $X(x) \geq \max\{S_j(x) : 1 \leq j \leq n\} - M_n(T(x))$. Also, notice that we have specifically defined $S_0(y) = 0$ for all $y \in \Omega$. So, $S_0(T(x)) = 0$. Since our functions M_n are defined as the maximum values of all the functions S_j, including S_0, then we know that $0 \in \{S_j(x) : 0 \leq j \leq n\}$ and so $M_n(T(x)) \geq 0$.

Next, since $M_n(x) > 0$, then we know that we do not need to include the $S_0(x) = 0$ element when computing $M_n(x)$. Therefore, $M_n(x) = \max\{S_j(x) : 1 \leq j \leq n\}$. So, recall that we just showed that $X(x) \geq \max\{S_j(x) : 1 \leq j \leq n\} - M_n(T(x))$. We now have that $X(x) \geq M_n(x) - M_n(T(x))$ for all x such that $M_n(x) > 0$. We also know that $M_n(x)$ is just a maximum value of functions that are summations of the integrable functions X. Therefore, by the properties of the integral, since

$X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$, $M_n \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$, as well. This means that we can happily take its expectation value. So, using the inequalities we have proven above, we find that

$$\begin{aligned}
&E[X(x)I_{\{y\in\Omega\,:\,M_n(y)>0\}}(x)] \geq E[(M_n(x) - M_n(T(x)))I_{\{y\in\Omega\,:\,M_n(y)>0\}}(x)] \\
&\geq E[M_n(x) - M_n(T(x))I_{\{y\in\Omega\,:\,M_n(y)>0\}}(x)] \\
&\geq E[M_n(x) - M_n(T(x))I_{\{y\in\Omega\,:\,M_n(y)\geq 0\}}(x)] \\
&= E(M_n(x) - M_n(T(x))) \\
&= 0
\end{aligned}$$

To be a little more specific, the first line of the inequalities follows from the fact that $X(x) \geq M_n(x) - M_n(T(x))$ for all x such that $M_n(x) > 0$ (what we just showed). The second line follows because the indicator function forces our integrand to be non-zero only in the places where $M_n(x) > 0$. Hence, when we allow the term $M_n(x)I_{\{y\in\Omega\,:\,M_n(y)>0\}}(x)$ to become just $M_n(x)$ in the integrand, we can only be contributing terms that are less than or equal to zero. Consequently, we have the inequality in the second line. The third inequality comes from the fact that $M_n(T(x)) \geq 0$. Thus, since the set $\{y \in \Omega\,:\,M_n(y) > 0\} \subseteq \{y \in \Omega\,:\,M_n(y) \geq 0\}$, then we know that changing the indicator function as shown in the third inequality can only make the integral smaller because we are subtracting the product of the positive functions $M_n(T(x)))I_{\{y\in\Omega\,:\,M_n(y)\geq 0\}}$. The equality in the fourth line simply follows from the fact that $M_n(x) \geq 0$ for all $x \in \Omega$. Again, this is true because we always include $S_0(x) = 0$ when taking that maximum in the definition of $M_n(x)$. Finally, the very last equality follows from the fact that T is measure preserving. Indeed, since the measure is not changing the probability structure of our space Ω, then the expectation of any random variable X defined on Ω is the same as the expectation of a variable on $T(\Omega)$. We are now done with the proof. □

We may now finally move on to the proof of Birkhoff's Ergodic Theorem. It is outlined on page 275 in Athreya's book.

Proof. We will assume that our expected value $E(X)$ is equal to zero. This can be done without a loss of generality because we know that $E(X)$ is some finite number due to our assumption that $X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$. Therefore, we can arbitrarily change the value of $E(X)$ by shifting our random variable function by any $\alpha \in \mathbb{R}$. In other words, we can always just think about $X'(x) = X(x) + \alpha$. So, since shifts like this do not change our arguments, we can simply say that $E(X) = 0$ to make the notation a little easier. Anyway, let's begin. We again want to consider summations of our random variable functions. So, just as in Lem. 5.2.2, let us define $S_n(x) = \sum_{j=0}^{n-1} X(T^j(x))$ for each

$n \geq 1$ and for all $x \in \Omega$. Also, $S_0(x) = 0$ for all $x \in \Omega$. Now, we will also consider the function $Z(x) \equiv \limsup_{n\to\infty} S_n(x)/n$. Notice that the function $Z(x)$ is just the supremum limit of the sequence of functions (f_n) in the statement of the ergodic theorem. Anyway, let us fix an arbitrary tolerance $\epsilon > 0$ and look at the set $A_\epsilon \equiv \{x \in \Omega : Z(x) > \epsilon\}$. So, what we are doing here is looking at the set of all sample points in our sample space that are initial conditions that yield time averages (the $Z(x)$'s) that are "ϵ-far away" from our expected value $E(X) = 0$. Anyway, we want to show that $\mu(A_\epsilon) = 0$. In other words, we want to show that the set of all such initial conditions has a measure zero. Indeed, we want to make sure that almost every initial condition converges yields a time average of X that converges to the expected value. Now, it is important to note here that A_ϵ is T-invariant. In other words, $A_\epsilon = T^{-1}(A_\epsilon)$. We will show this now.

($\subseteq$): Suppose that $y \in A_\epsilon$. To show that $y \in T^{-1}(A_\epsilon)$, we just have to prove that $T(y) \in A_\epsilon$. So, $y \in A_\epsilon$ means that $Z(y) > \epsilon$. Therefore, $\limsup_{n\to\infty} S_n(y)/n > \epsilon$. However, we know that $S_n(y) = \sum_{j=0}^{n-1} X(T^j(y))$ for all $n \geq 1$. Therefore, $S_n(T(y)) = \sum_{j=0}^{n-1} X(T^{j+1}(y)) = S_{n+1}(y) - S_0(y) = S_{n+1}(y)$. Moreover, $S_{n+1}(y) = S_n(y) + X(T^n(y))$. We may now reason that

$$\limsup_{n\to\infty} \frac{S_n(T(y))}{n} = \limsup_{n\to\infty} \left(\frac{S_n(y)}{n} + \frac{X(T^n(y))}{n} \right) = \limsup_{n\to\infty} \frac{S_n(y)}{n} > \epsilon.$$

The last equality follows from the fact that $X(T^n(y))$ must be bounded because we know that $X \in \mathcal{L}(\Omega, \mathcal{F}, \mu)$, and certainly we cannot get finite integrals of unbounded functions. Anyway, $\frac{X(T^n(y))}{n} \to 0$ as $n \to \infty$. We have now shown that $T(y) \in A_\epsilon$.

($\supseteq$): Suppose that $y \in T^{-1}(A_\epsilon)$. This means that $T(y) \in A_\epsilon$. We just have to show that $y \in A_\epsilon$. Again, $T(y) \in A_\epsilon$ means that $\limsup_{n\to\infty} S_n(T(y))/n > \epsilon$. However, as we saw before, this lim sup is just equal to $\limsup_{n\to\infty} S_n(y)/n$. Thus, $y \in A_\epsilon$, as well.

We have now shown that $T^{-1}(A_\epsilon) = A_\epsilon$. Of course, in the hypothesis of our theorem, we assumed that T is ergodic. Thus, we know that either $\mu(A_\epsilon) = 1$ or $\mu(A_\epsilon) = 0$. Since we want to show the latter condition, we will assume that $\mu(A_\epsilon) = 1$ and look for a contradiction. Now, we will define another random variable $Y(x) = X(x) - \epsilon$. Similarly to the Lem. 5.2.2, we define the functions $M_{n,Y}(x) \equiv \max\{S_{j,Y}(x) : 0 \leq j \leq n\}$, where we now have $S_{0,Y}(x) \equiv 0$ and $S_{j,Y}(x) \equiv \sum_{i=0}^{j-1} Y(T^k(x))$ for all $j \in \mathbb{N}$. These sequences of functions are identical to the ones in the conditions for Lem. 5.2.2, so we conclude by that lemma that

$$E(Y(x) I_{\{y\in\Omega\,:\, M_{n,Y}(y)>0\}}(x)) \geq 0.$$

From the definition of our functions $M_{n,Y}$, if we look at the sets $B_n \equiv \{x \in \Omega : M_{n,Y}(y) > 0\} =$

$\{x \in \Omega : \sup_{1\leq j\leq n}\{\frac{1}{j}S_{j,Y}(x)\} > 0\}$, we find that $B_n \uparrow B \equiv \{x \in \Omega : \sup_{1\leq j\leq\infty}\{\frac{1}{j}S_{j,Y}(x)\} > 0\}$. We see that we have an $\uparrow$ symbol here because $B_n \subseteq B_{n+1}$ by definition since each time we increase our index n, we add one more element to the supremum we are computing in the definition of B_n. Thus, whatever elements $x \in \Omega$ were in B_n, we must have the same elements in B_{n+1}, since our supremum value can only increase. In other words, once $\sup_{1\leq j\leq n}\{\frac{1}{j}S_{j,Y}(x)\} > 0$ for some $x \in B_n$, then certainly $\sup_{1\leq j\leq n+1}\{\frac{1}{j}S_{j,Y}(x)\} > 0$ for that same x. So, $x \in B_{n+1}$. Then, we can always take the countable union of these sets to get some limiting set B. This is what we mean by the statement $B_n \uparrow B$. Now let us consider some $z \in A_\epsilon$. This means that $Z(z) > \epsilon$. However, from our definition of Y, $\frac{1}{j}S_{j,Y}(x) = \frac{1}{j}S_j(x) - \epsilon$ for $j \geq 1$. Therefore,

$$Z(z) = \limsup_{n\to\infty}\left(\frac{S_n(z)}{n}\right) = \limsup_{n\to\infty}\left(\frac{1}{n}S_{j,Y}(z) + \epsilon\right) > \epsilon \;\Rightarrow\; \limsup_{n\to\infty}\left(\frac{1}{n}S_{j,Y}(z)\right) > 0.$$

We now conclude that $\sup_{1\geq j<\infty}\frac{1}{j}S_{j,Y}(z) > 0$ and so $z \in B$. Consequently, $A_\epsilon \subset B$. However, we assumed at the beginning that $\mu(A_\epsilon) = 1$. Hence, $\mu(B) = 1$, as well. Finally, since X is integrable, $|Y| = |X - \epsilon| \leq |X| + \epsilon$ is integrable, as well. Thus, we can construct a sequence of measurable functions by looking at $Y\,I_{B_n}$. Then, since $B_n \uparrow B$, we certainly have the case that the sequence of functions $(Y\,I_{B_n})$ converges pointwise to $(Y\,I_B)$ Furthermore, $\int_\Omega Y I_B\,d\mu = \int_\Omega Y\,d\mu$ because $\mu(B) = 1$ and so B is differs from Ω by a set of measure 0 (i. e. $\mu(\Omega \setminus B) = 0$). This means that $YI_B = Y$ almost everywhere in Ω, and so our integral values are the same. Anyway, since we also have the bound $|YI_{B_n}| \leq |Y| \leq |X| + \epsilon$, and $|X| + \epsilon$ is certainly integrable, then we can use Lebesgue's dominated convergence theorem to conclude that

$$E(YI_{B_n}) = \int_\Omega YI_{B_n}\,d\mu \;\longrightarrow\; \int_\Omega YI_B\,d\mu = E(YI_B) = E(Y) \text{ as } n \to \infty.$$

However, this is a contradiction because we have just shown that $E(YI_B) = E(YI_{\{y\in\Omega\,:\,M_{n,Y}(y)>0\}}) = E(Y) = -\epsilon < 0$. Thus, we are contradicting the result we got from Lem. 5.2.2. Anyway, we now know that $\mu(A_\epsilon) = 0$. We see that this must be the case since having $\mu(A_\epsilon) = 0$ prevents us from inferring that $\mu(B) = 1$ and reaching a contradiction. Indeed, Lem. 5.2.2 necessitates the result that $\mu(A_\epsilon) = 0$. To finish our argument, notice that $A_\epsilon^{\mathcal{C}} = \{x \in \Omega : Z(x) \leq \epsilon\}$. Also, we necessarily have that $\mu(A_\epsilon^{\mathcal{C}}) = 1$ since $\mu(A_\epsilon) = 0$. However, $\mu(A_\epsilon) = 0$ is true for every $\epsilon > 0$ because our original $\epsilon > 0$ was arbitrary. It follows that our condition $Z(x) \leq \epsilon$ in the definition of $A_\epsilon^{\mathcal{C}}$ becomes just $Z(x) \leq 0$ from elementary real analysis considerations. Therefore, given our definition of Z, we have just shown that

$$\mu\left(\left\{x \in \Omega : \limsup_{n\to\infty}\frac{S_n(x)}{n} \leq 0\right\}\right) = 1.$$

We can go through *a completely parallel argument* for the random variable $-X$. In this case, all of the inequalities become flipped and we have lim inf's instead of lim sup's. Therefore, by the symmetry of the argument above, we also see that

$$\mu\left(\left\{x \in \Omega \,:\, \liminf_{n\to\infty} \frac{S_n(x)}{n} \geq 0\right\}\right) = 1.$$

However, we know that for any given sequence of real numbers (a_n), $\liminf a_n \leq \limsup a_n$. Further, we have just shown that $0 \leq \liminf \frac{S_n(x)}{n}$ and $\limsup \frac{S_n(x)}{n} \leq 0$ for almost every $x \in \Omega$. So, the supremum limit and infimum limit have now been squeezed together to the point where they both must equal zero (almost everywhere)! This means that our sequence converges! Indeed, we have that

$$\mu\left(\left\{x \in \Omega \,:\, \lim_{n\to\infty} \frac{S_n(x)}{n} = 0 = E(X)\right\}\right) = 1.$$

We have now shown the first part of the theorem, since $S_n(x)/n$ is just a short-hand notation for our time average. Therefore, we have shown that as $n \to \infty$, the probability that our time average is equal to the phase space average (which we set to zero in the proof) goes to one! This is precisely the definition of convergence *with probability* 1. This is great, since we are almost done.

The last bit of the proof simply deals with convergence of the time average to the expected value in $\mathcal{L}$. To remind ourselves of what this means, recall Def. 3.5.3. So, we just want to show that the following sequence of integrals (which is just a sequence of real numbers) converges: $\int_\Omega |\frac{1}{n}\sum_{j=1}^{n-1} X(T^j(x))|\, d\mu \to 0 = E(X)$ as $n \to \infty$. In order to do this, we can split up our random variable X into its positive, X^+, and negative, X^-, parts. Then, we can make the *exact same arguments* we made above for X to show that the time averages of X^+ and X^- converge to $E(X^+)$ and $E(X^-)$, respectively, with probability 1. However, consider our time average of either $X^\pm$. We know that it is written as $\frac{1}{n}\sum_{i=0}^{n-1} X^\pm(T^i(x))$ for any initial point $x \in \Omega$. Since T is measure preserving, $\int_\Omega X^\pm(T^i(x))\, d\mu = \int_\Omega X^\pm(x)\, d\mu$ for all $i \in \mathbb{N}$. Therefore, the integral of our time average is just $\int_\Omega \frac{1}{n}\sum_{i=0}^{n-1} X^\pm(T^i(x))\, d\mu = \frac{1}{n}\sum_{i=0}^{n-1} \int_\Omega X^\pm(x)\, d\mu = \int_\Omega X^\pm(x)\, d\mu = E(X^\pm)$. Clearly this must be true for all $n \in \mathbb{N}$. Anyway, notice that we now have two sequences $(\frac{1}{n}\sum_{i=0}^{n-1} X^+(T^i(x)))$ and $(\frac{1}{n}\sum_{i=0}^{n-1} X^-(T^i(x)))$ of nonnegative (since X^- and X^+ are defined as nonnegative) measurable functions that converge pointwise to $E(X^+(x))$ and $E(X^-(x))$, respectively. We also found that $\int_\Omega \frac{1}{n}\sum_{i=0}^{n-1} X^\pm(T^i(x))\, d\mu \to \int_\Omega E(X^\pm)\, d\mu = E(X^\pm)$. Thus, we have all the necessary conditions to apply Scheffe's Theorem (Thm. 3.4.7). Applying it we get that

$$\int_\Omega \left|\frac{1}{n}\sum_{i=0}^{n-1} X^\pm(T^i(x)) - E(X^\pm(x))\right| d\mu \to 0$$

as $n \to \infty$. Indeed, since we have now shown that this convergence in $\mathcal{L}(\Omega, \mathcal{F}, \mu)$ is valid for the two parts X^+ and X^-, since $X = X^+ - X^-$, we conclude that as $n \to \infty$

$$\lim_{n\to\infty} \int_\Omega \left| \frac{1}{n} \sum_{i=0}^{n-1} X(T^i(x)) - E(X) \right| d\mu \to 0.$$

We have now shown the whole theorem! This last bit is just the definition of convergence in $\mathcal{L}(\Omega, \mathcal{F}, \mu)$. □

5.3 Conclusions

The groundwork for ergodic theory is now complete. We are ready to tackle further interesting properties of measure-preserving transformations on measure spaces, such as mixing and entropy. Walter's book [Wa] provides an advanced introduction to these topics. However, for now, we will be best served by pausing and reflecting on the importance of what we have already established. Birkhoff's ergodic theorem is a uniquely powerful result that has far-reaching implications. To conclude our discussion, we will go over three areas which have been influenced by this theorem.

First, from the point of view of probability theory, the ergodic theorem is a powerful *strong law of large numbers* (SLLN). Specifically, the most common form of the SLLN may be phrased as a theorem (pg. 240 [Ath]):

Theorem 5.3.1. *Let (X_n) be a sequence of independent and identically distributed random variables such that $E(X_1^4) < \infty$. Then,*

$$\bar{X_n} \equiv \frac{X_1 + X_2 + \ldots + X_n}{n} \to E(X_1)$$

with probability 1 as $n \to \infty$. The convergence above is a pointwise convergence of the sequence of functions $\bar{X_n}$ to the constant function $E(X_1)$.

Notice that the language of the SLLN theorem given above is very similar to the language of Birkhoff's ergodic theorem. In both theorems, a sequence of finite sums of random variables converges with probability 1 to the expectation of the random variable. In the SLLN case, we have many independent, identically distributed random variables. In the ergodic theorem, we are concerned with a single random variable whose domain Ω is being acted on by the ergodic transformation T. Thus, although

neither theorem implies the other directly, we see that the ergodic theorem has the advantage of not requiring independence or identical distributions. This is why we may refer to the latter theorem as an especially strong law of large numbers.

A second field that has been impacted by the ergodic theorem is number theory. Specifically, one is able to prove "Borel's Theorem on Normal Numbers" as an almost direct corollary of the ergodic theorem (pg. 35 in [Wa]). This is a fascinating result that states that for almost every $x \in [0,1)$, the frequency of 1's (and 0's) in the binary expansion of x is $1/2$! This means that except for some subset of measure zero, all the real numbers in the interval $[0,1)$ have approximately the same number of zeros and ones in their binary expansions! Although we will not go through the proof of this result, we will present the arguments in very broad terms. The basic idea is that one can construct an ergodic map $T(x) = 2x \mod 1$ where $T : [0,1) \to [0,1)$ (see the example of the measure-preserving transformation on $[0,1]$ in Sec. 5.1 for the explicit definition of T). This map shifts the binary expansion of any given element in $[0,1)$ by one number. In other words, if our binary expansion of some $x \in [0,1)$ is $x = 1/2 + 0/4 + 1/8 + 1/16 + 0/32\ldots$, then $T(x) = 0/2 + 1/4 + 1/8 + 0/16 + \ldots$. Next, one constructs the random variable $X(x) = I_{[1/2,1)}(x)$ on $[0,1)$. Notice that $X(x)$ is 1 when the leading term in the binary expansion of x is 1, and 0 otherwise. Therefore, by combining this random variable (measurable function) with the shifting property of the T transformation, we find that $\sum_{i=0}^{n-1} X(T^i(x))$ represents the number of 1's in the first n digits of the binary expansion! An application of the ergodic theorem yields the remarkable result. For a more explicit proof, we refer the reader to pages 35 and 36 in [Bu].

The third and final area which we will mention is statistical mechanics. The relationship between this field and ergodic theory cannot be overstated since the concept of ergodicity itself originated in this field. Indeed, the word "ergodic" was first coined by one of the founders of statistical mechanics, Ludwig Boltzmann. It comes from the Greek word "ergon", meaning work, and "odos", meaning path. So, the start of ergodic theory began with Boltzmann's hypothesis that the orbit of a transformation T on a phase space Ω (a set of all possible states of a system) would yield the entire phase space (pg. 1,2 in [Wa]).

In statistical mechanics, the transformation T represents the evolution of a system through its possible states. It is, in a sense, a *process*. For example, if the system is at a point x in its phase space at some time t_0, then $T_t(x)$ may represent the state of the system at time $t_0 + t$. Anyway, Boltzmann expected that the time and phase space averages would be the same for these physical systems (the two averages discussed in the Sec. 5.2). Such a condition is valuable in statistical mechanics

because one expects that a physical system will sample all of the states available to it. This is in fact the pillar on which most of statistical mechanics is built. Of course, we now know that such an assumption is justified by Birkhoff's ergodic theorem only for specific types of transformations which we call ergodic. Physicists in modern times always have to consider that ergodicity is not a trivial condition satisfied by most physical processes, but rather a deep mathematical property of special transformations (processes) on measure spaces (states of a system) with far reaching consequences.

6 Bibliography

[Ad] Adams, M. and Guillemin, V. (1996), *Measure Theory and Probability*, Birkhäuser, Boston

[As] Asplund, E. and Bungart, L. (1966), *A First Course in Integration*, Holt, Rinehart, and Winston, New York

[Ath] Athreya, K. B. and Lahiri, S. N. (2006), *Measure Theory and Probability Theory*, Springer, New York.

[Bi] Birkhoff, G. D. (1931), 'Proof of a Recurrence Theorem for Strongly Transitive Systems' and 'Proof of the Ergodic Theorem', *Proc. of the Nat. Acad. of Sciences* **17**(12), 650-660

[Bo] Bogachev, V. I. (2000), *Measure Theory*, Springer, New York

[Bu] Burk, F. (1998), *Lebesgue Measure and Integration: An Introduction*, John Wiley & Sons, New York

[Sc] Schumacher, C. S. (2007), *Closer and Closer: Introducing Real Analysis*, Jones & Bartlett Pub., Sudbury, Massachusetts

[Wa] Walters, P. (1982), *An Introduction to Ergodic Theory*, Springer-Verlag, New York

[Wh] Wheeden, R. L. and Zygmund, A. (1977), *Measure and Integral: An Introduction to Real Analysis*, Marcel Dekker, New York

[Wi] Image provided by the Wikipedia article: "Cantor set" (*http://en.wikipedia.org/wiki/Cantorset*, page last modified on October 27, 2007)