Stochastic Techniques in Concurrency

Prakash Panangaden
School of Computer Science
McGill University
Montréal, Québec

Chapter 1

Introduction

1.1 Preliminary Remarks

1.2 The Need For Measure Theory

Elementary probability theory can be summed up easily. Imagine that one has a process which makes a single step and can end up in any one of a finite set S of final states each with equal likelihood. Then the probability than the final state lies in a subset A - often called an **event** - is given by $|A|/|S|$ where $|\cdot|$ denotes the size of a finite set. From this simple intuition one can define concepts like the probability of more complex processes which might involve several steps or interaction between different observations.

The typical concepts that one learns: independence, expectation value, and conditional probability are fairly clear - at least in their intuitive conception - in the "discrete" case described above. These concepts suffice to analyze much of the work in probabilistic process algebra. In some sense the relevant concepts are essentially those of Boolean algebra. However, in the continuous case, the same concepts require different mathematics. In some sense one can say that one has to move from Boolean algebras to σ-Boolean algebras. Measure theory evolved originally to make sense of integration but - essentially in Kolmogorov's hands - it became the rigourous foundation for probability theory. The need for such extensions to high-school probability theory arose from statistical mechanics and the need to explain physical phenomena like Brownian motion.

For researchers interested in systems like process control systems, telecommunication systems, networks there are very similar phenomena. There is a uncontrolled physical phenomenon, "noise" or "drift" and some controlling software. Understanding how these interact is essential for the design and analysis of such systems.

In order to see how measure theory is forced we will consider a classical example - an infinite sequence of coin tosses. This is paradigmatic of an infinitely repeated operation and will be relevant for any analysis of recursive or indefinite iteration in a probabilistic setting. Even though the basic actions are discrete we are led to measure theory by the infinite repetition. Now if we asked naive questions such as "what is the probability of the sequence $(HT)^{\infty}$" we would get 0 as the answer. From this alone we can conclude very little. Right away we observe a striking difference from the finite case. Knowing all the singleton probabilities does not tell us the probabilities associated with other sets. The singleton sets are no longer the "atomic building blocks" from which everything else can be built. We want to be able to say things like "the probability of getting infinitely many heads is 1" which we certainly cannot conclude by simple counting arguments.

What we need is a suitable notion which allows us to define the probabilities in a suitable limiting fashion. We expect that there are certain sets that we can easily associate probabilities to and such that we can define the probabilities associated with other sets by operations. But what are the reasonable operations? It seems compelling that the operations of the discrete theory should survive - these are finite union, finite intersection and complementation. Thus we expect that we will have a family of sets closed under these operations. We further expect that

$$Pr(A) \wedge Pr(B) = Pr(A \cap B)$$

with similar formulas for disjoint union and complementation. We have seen that we cannot expect a summation formula for *arbitrary* unions but, if we want limits to be computable, we can demand that *countable* unions behave like finite unions. In other words we demand that the family of sets that we are working with be closed under countable union and complement; intersection is, of course, superfluous. We demand that if we have a pairwise disjoint family of sets A_i then

$$Pr(\cup_i A_i) = \sum_i Pr(A_i) \text{ and } Pr(A^c) = 1 - P(A).$$

From this can we compute the probability of having infinitely many heads? The probability of having the first toss be a head followed by an infinite sequence of tails is 0. The probability of exactly one head anywhere is again 0, by considering the countable union. The probability of any fixed finite number of heads is 0, again by taking a countable union and the probability of finitely many heads is again 0. Thus the probability of infinitely many heads is 1. Of course not all answers should be 0 and 1. The probability of a head followed by an *arbitrary* sequence should be 1/2. The sets which look like initial finite sequences followed by an arbitrary sequence are the sets which serve as the basis from which to compute all probabilities.

This raises the natural questions: can we compute probabilities for all the sets this way? It turns out that the answer is "no"! There are sets for which probability or "measure" cannot be sensibly defined. This never happens when the space of outcomes or states is countable but happens in "almost" any uncountable space.

The key point to take away from this is that we expect to work with *countable* operations - finite ones are not enough and arbitrary ones are impossible.

1.3 The Laws of Large Numbers

One of the most striking early results in probability theory was Borel's law of large numbers. In fact there are two: the **strong** law and the **weak** law. It is worth understanding what they say to see what the difference between discrete and continuous probabilities is. The weak law can be explained (and proved) entirely in terms of discrete probability, but the **strong** law requires the ideas of continuous probability distributions. One can think of this law as corresponding to the choice of a real number from the unit interval or as an infinite sequence of coin tosses. By viewing a sequence of heads and tails as the binary encoding of a real number we can relate the two[1]. We will work with the coin tossing interpretation.

[1] Modulo some minor niggling details about representing numbers uniquely.

We imagine that we have a fair coin - it has equal probabilities of producing a head or a tail - and we toss it infinitely often. We have seen some of the questions that we can ask about this situation, now we consider more subtle questions connected with deviations. Intuitively, we expect that the proportion of heads and tails should be half each, at least in some limiting sense. How could we formalize this? First we could say that the probability that there are n heads in the first $2n$ tosses tends to $1/2$ as n tends to 0. It can be stated in the following way. Instead of heads and tails we work with ones and zeros. We write s for a sequence of coin tosses and we write s_i for the value (1 or 0) of the ith toss. Then the number of heads (= 1s) in the first n tosses can be written as $\sum_{i=1}^{n} s_i$. An elementary argument shows that

$$Pr\{s : \sum_{i=1}^{n} s_i = k\} = \left(\begin{array}{c} n \\ k \end{array} \right) \frac{1}{2^n}.$$

The weak law expresses the intuition in terms of deviations from the expected probability.

$$\lim_{n \longrightarrow \infty} Pr\{s : |\frac{1}{n} \sum_{i=1}^{n} s_i - \frac{1}{2}| \geq \epsilon\} = 0$$

for any $\epsilon \geq 0$. This law is expressed entirely in terms of concepts that arise in discrete probability even though it says something nontrivial about a non-discrete situation. The proof can be found in many books, Billingsley [Bil95] or Breiman [Bre68] are good places to look.

Proving the Weak Law We include the proof for the interested reader. For the present paragraph we assume that the reader if familiar with the concept of expectation value of a random variable and of indicator or characteristic function of a set. These are explained later in the text. We proceed as follows. First we prove Chebyshev's inequality for the case of the space of sequences. Let Ω_n be the space of sequences of length n of heads and tails.

Proposition 1.1 (Chebyshev) *Let X be a function from Ω_n to the reals. Let $\mathsf{E}[X]$ stand for the expectation value of X.*

$$\forall \epsilon > 0. Pr(\{s : X(s) \geq \epsilon\}) \leq \frac{\mathsf{E}[X^2]}{\epsilon^2}.$$

Proof.

$$\begin{array}{ll} & Pr(\{s : X(s) \geq \epsilon\}) \\ = & \mathsf{E}[1_{\{s:X(s)\geq\epsilon\}}] \\ \leq & \mathsf{E}[\frac{X^2}{\epsilon^2} 1_{\{s:X(s)\geq\epsilon\}}] \\ \leq & \mathsf{E}[\frac{X^2}{\epsilon^2}] \\ = & \frac{1}{\epsilon^2}\mathsf{E}[X^2] \end{array}$$

■

ProofOf the weak law. We define the function $X_j : \Omega_n \longrightarrow \mathbf{R}$ by $X_j(s) = 1$ if the jth element of the sequence s is a head and 0 if it is a tail. We define the function $S_n : \Omega_n \longrightarrow \mathbf{R}$ by $S_n(s) := \sum_{i=1}^{n} X_i(s)$. Thus S_n counts the number of heads in s. Now we can proceed as follows. Note that by applying Chebyshev's inequality with the function $\frac{1}{n}S_n - \frac{1}{2}$ we have immediately

$$Pr(\{s : |\frac{1}{n}S_n - \frac{1}{2}| \geq \epsilon\}) \leq \mathsf{E}[|\tfrac{1}{n}S_n - \tfrac{1}{2}|^2]/\epsilon^2.$$

We can write $\frac{1}{n}S_n - \frac{1}{2}$ as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{2})$. Thus the expectation value that we need is $\frac{1}{n^2}\mathsf{E}[(\sum_{i=1}^{n}(X_i - \frac{1}{2}))^2]$. To calculate the expectation value we first note that if i and j are not equal then $\mathsf{E}[(X_i - \frac{1}{2})(X_j - \frac{1}{2})] = 0$. Thus when we square the sum all the cross terms have zero expectation. Thus we are left with $\mathsf{E}[\sum_{i=1}^{n}(X_i - \frac{1}{2})^2]$; each term of this sum has expectation $\frac{1}{4}$ so the sum has expectation $\frac{n}{4}$. Thus

$$Pr(\{s : |\frac{1}{n}S_n - \frac{1}{2}| \geq \epsilon\}) \leq \frac{1}{4n\epsilon^2}.$$

Now when we take the limit $n \longrightarrow \infty$ we get the result. ∎

Notice that this proof involves purely finite quantities and discrete probability theory.

The so-called *strong* law states something about the probability of sequences that satisfy a condition stated in terms of the entire infinite sequence. We define the set of interest as follows

$$A = \{s : \lim_{n \longrightarrow \infty}[\frac{1}{n}\sum_{i=1}^{n} s_i] = \frac{1}{2}\}.$$

This is the set of sequences with asymptotically equal number of heads and tails. Note that the condition of asymptotic equality has to be satisfied by every sequence in the set. When we view these sequences as numbers we get the set of numbers with equal occurrences of the two bits 1 and 0 in their binary representation. Such numbers are called *normal* numbers base 2. One can similarly define normal number for any base. A *normal number* is normal with respect to any base. Borel's normal number theorem says that the probability of choosing a non-normal number at random is 0. This is a statement that makes no sense unless we have a notion of *measure* on the entire sets of infinite sequences.

The strong law implies the weak law but not vice-versa. There are two different notions of convergence at work here. In the weak law we have convergence of the expected values whereas in the strong law we have exact convergence holding with probability 1; what is often called *almost sure* convergence.

1.4 Borel-Cantelli Lemmas

In the last section we have been considering situations where we have to talk about infinitely many occurrences. In this section we describe two classical lemmas about this type of situation. Suppose that we have a situation, say discrete for simplicity, where there are infinitely many events of interest. We write $\{A_i : i \in \mathbf{N}\}$ for these events. The situation that we have in mind is the following. We repeat the experiment infinitely often. In each repetition one (or more) of the A_i may occur or perhaps none of them occur. Now how do we describe the situation "the A_i happen infinitely often" in repeated trials? We are really looking at the countable product space of sequences of trials as we did with the discussion of infinite sequences of heads and tails. The **lim sup** of the sets A_i is given by

$$A = \limsup_{n \longrightarrow \infty} A_n = \bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty} A_n.$$

This is the set theoretical analogue of the "lim sup" which you might remember from undergraduate analysis. It corresponds to the situation we wish to describe, i.e. infinitely often one of the A_i happens. We can now state the first Borel-Cantelli lemma.

Proposition 1.2 (Borel-Cantelli I) *With the notation above, if* $\sum_{n=1}^{\infty} Pr(A_i) < \infty$ *then* $Pr(A) = 0$.

Proof. Let $B_N = \cup_{n=N}^{\infty} A_n$ then $A = \cap_{N=1}^{\infty} B_N$. Now we have $\forall N. A \subseteq B_N$, thus

$$Pr(A) \leq Pr(B_N) \leq \sum_{n=N}^{\infty} Pr(A_n).$$

Now if we let N go to infinity the sum on the right hand side must go to 0 since it is the tail of a convergent sequence, recall we have $\sum_{n=1}^{\infty} Pr(A_n) < \infty$. Thus $Pr(A) = 0$. ∎

This makes no assumption about the events being independent. The converse result does, however, require an independence hypothesis.

Proposition 1.3 (Borel-Cantelli II) *If the* A_n *are independent then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \text{ implies } P(A) = 1.$$

Proof. Let us write C_n for the complement of A_n, thus $P(C_n) = 1 - P(A_n)$. Note the obvious inequality $(1 - x) < e^{-x}$ for any x in $(0, 1)$. For any N and $M > N$ we have:

$$\begin{aligned}
& Pr(\textstyle\bigcap_{n=N}^{\infty} C_n) \\
\leq\ & Pr(\textstyle\bigcap_{n=N}^{M} C_n) \\
& \text{now, since we have assumed independence we have,} \\
=\ & \textstyle\prod_{n=N}^{M} Pr(C_n) \\
=\ & \textstyle\prod_{n=N}^{M} (1 - Pr(A_n)) \\
\leq\ & \textstyle\prod_{n=N}^{M} \exp(-Pr(A_n)) \\
=\ & \exp(-\textstyle\sum_{n=N}^{M} Pr(A_n)).
\end{aligned}$$

We used independence in the first equality above. Now if we let M go to ∞ the rhs goes to 0 since the exponent goes to $-\infty$ by hypothesis. Now we have

$$A^c = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} C_n$$

hence

$$Pr(A^c) \leq \sum_{N=1}^{\infty} Pr(\bigcap_{n=N}^{\infty} C_n) = 0.$$

Thus we have $Pr(A) = 1$. ∎

These lemmas are prototypes for what are called “zero-one” laws. If we assume independence we have the result that the probability of A is always either 1 or 0. Consider our coin-tossing example. If the probability of heads is some fixed number greater than 0 the probability that we will have a sequence of 1729 consecutive heads is 1! Thus even though we expect the numbers to average out “in the long run”, in any given sequence of repeated experiments we expect, with high probability, that there are arbitrary long sequences of heads.

Chapter 2

Measure Theory

In this chapter we discuss the axioms for measure theory from an abstract point of view. This will prepare us for both major applications of the ideas, viz to probability theory and to integration. Intuitively a measure is a notion of "size" that one wishes to attach to sets. This notion is intended to reflect the geometric notion of size coming from examples like area and volume. Measure turns out to be poorly related to set theoretic conceptions of size. It would be pleasant if we could take all sets to be measurable; unfortunately, as we shall see below, this is not possible, even for such common spaces as $\mathbf{R}$. In situations with a countable set of possible states we can indeed take all sets to be measurable and much of the subtleties of measure theory can be dispensed with. However, results and proofs obtained in the discrete case are not a very reliable guide to the continuous case.

In the present chapter we have given most proofs in greater detail than in most books and we have assumed little prior exposure to advanced analysis. We do assume basic notions in topology and metric spaces. Furthermore we assume some basic terminology from category theory. In writing this chapter I have benefited greatly from the books of Rudin [Rud66], Geroch [Ger85], Kingman and Taylor [KT66] and Malliavin [Mal95].

2.1 Measurable Spaces

Definition 2.1 *A* **measurable space** (X, Σ) *is a set* X *together with a family of subsets of* X*, called a* σ**-field** *or* σ**-algebra***, satisfying the following axioms:*

1. $\emptyset \in \Sigma$,
2. $A \in \Sigma$ *implies that* $A^c \in \Sigma$*, and*
3. *if* $\{A_i \in \Sigma | i \in I\}$ *is a countable family then* $\cup_{i \in I} A_i \in \Sigma$.

If we require only finite additivity rather than countable additivity we get a ***field***.

Notice that, unlike open sets in a topology, measurable sets are closed under complementation and hence under countable intersections as well. This makes a dramatic difference to the properties of measurable functions, compared with continuous functions, as we shall see below. Note also that singletons may or may not belong to a σ-field. In most σ-fields that we are interested in the singletons will be measurable sets.
We develop some of the basic properties of σ-fields.

Proposition 2.2 *The intersection of an arbitrary collection of σ-fields on a set X is a σ-field on X.*

The proof is left as an exercise. The following corollary is very important.

Corollary 2.3 *Given any subset $\mathcal{B}$ of $\mathcal{P}(X)$ there is a least σ-field containing $\mathcal{B}$.*

We often refer to the least σ-field containing $\mathcal{B}$ as the σ-field **generated** by $\mathcal{B}$.

Example 2.4 Given a set X the powerset $\mathcal{P}(X)$ is a σ-field. The set consisting of just X and $\emptyset$ is another σ-field. If X is a countable set and all singletons are measurable the σ-field is $\mathcal{P}(X)$. This is the situation with most discrete situations.

These are extreme examples of course. A more interesting example and a good source of counter-examples is the following.

Example 2.5 Let X be an uncountable set. The collection of all countable (finite or infinite) sets and cocountable sets (complements of countable sets) forms a σ-field on X. This is the σ-field generated by the singletons.

The next example is of fundamental importance.

Example 2.6 Given a topological space $(X, \mathcal{T})$ we define $\mathcal{B}(X)$ to be the σ-field generated by the open sets (or, equivalently, by the closed sets). Strictly speaking we should write $\mathcal{B}(X)$ since the σ-field depends on the topology and not just on the set X but it is customary to write as we have done since the topology is usually clear from context. The sets in the σ-field $\mathcal{B}(X)$ are called the **Borel sets**. The most important instance of this is the collection of Borel sets generated by the usual topology in $\mathbf{R}$. One often says "Borel sets" to refer to this special case.

The following propositions will be useful when we discuss measurable functions. The proofs are exercises.

Proposition 2.7 *Let $f : X \longrightarrow Y$ be a function and let Σ be a σ-field on Y. The sets of the form $\{f^{-1}(A) | A \in \Sigma\}$ form a σ-field on X.*

Quite often we want to work with the members of a set which generates a σ-field rather than with all sets in the σ-field.

Proposition 2.8 *Suppose that X, Y are sets and $f : X \longrightarrow Y$ is a function. Suppose that $\mathcal{G}$ is a family of subsets of Y and Σ is the σ-field generated by $\mathcal{G}$. Then $\{f^{-1}(A) | A \in \Sigma\}$ is the σ-field generated by $\{f^{-1}(A) | A \in \mathcal{G}\}$.*

The class of σ-fields can be characterized in terms of monotonicity properties; this will be very useful when we discuss integration. We introduce the following convenient notation. If we have a nested family of sets

$$A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \ldots$$

with $\cup_n A_n = A$ we write $A_n \uparrow A$. Similarly if

$$A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \ldots$$

with $\cap_n A_n = A$ we write $A_n \downarrow A$.

Definition 2.9 *A collection of sets $\mathcal{M}$ is called a* **monotone class** *if whenever $A_n \uparrow A$ with all $A_n \in \mathcal{M}$ then $A \in \mathcal{M}$ and also if $A_n \downarrow A$ with all $A_n \in \mathcal{M}$ then $A \in \mathcal{M}$.*

Clearly any σ-field is a monotone class and, just as for σ-fields, the intersection of monotone classes is a monotone class. Thus we can talk about the monotone class generated by a collection of sets just as we did for σ-fields. Recall that a field is like a σ-field except that we only require finite additivity rather than countable additivity. It is clear that a field that is also a monotone class must be a σ-field.

Proposition 2.10 *If $\mathcal{F}$ is a field of subsets of X then the monotone class, $\mathcal{M}$, generated by $\mathcal{F}$ is a σ-field on X.*

Proof. We define $\Omega(P)$ for any $P \subseteq X$ as follows

$$\Omega(P) = \{Q \subseteq X : P - Q, Q - P, P \cup Q \in \mathcal{M}\}.$$

Clearly $P \in \Omega(Q)$ iff $Q \in \Omega(P)$. Consider a sequence of sets $Q_1 \subseteq Q_2 \subseteq \ldots$ in $\Omega(P)$. We have $P \cup (\cup_n Q_n) = \cup_n (P \cup Q_n) \in \mathcal{M}$, hence $\cup_n Q_n \in \Omega(P)$. Similarly we can check that $P - (\cup_n Q_n)$ and $(\cup_n Q_n) - P$ are in $\Omega(P)$. Furthermore we can also show that $\Omega(P)$ is closed under countable intersection of nested families of sets in an analogous way. Thus each $\Omega(P)$ is a monotone class. Clearly $\mathcal{F} \subseteq \Omega(P)$ for any P in $\mathcal{F}$, hence $\mathcal{M} \subseteq \Omega(P)$ for any $P \in \mathcal{F}$. Now fix $Q \in \mathcal{M}$ and $P \in \mathcal{F}$ so that $Q \in \Omega(P)$, but then $P \in \Omega(Q)$ hence $\mathcal{F} \subseteq \Omega(Q)$ so $\mathcal{M} \subseteq \Omega(Q)$. This means that if we take P, Q in $\mathcal{M}$ then $P - Q$ and $P \cup Q$ are in $\mathcal{M}$. Thus we have in particular that $\mathcal{M}$ is closed under complementation and finite unions, i.e. it is a field and hence a σ-field. ∎

2.2 Measurable Functions

The above propositions show that σ-fields behave well under inverse images. Accordingly it is natural to define measurable functions in terms of inverse images.

Definition 2.11 *A function f from a σ-field (X, Σ_X) to a σ-field (Y, Σ_Y) is said to be* **measurable** *if $f^{-1}(B) \in \Sigma_X$ whenever $B \in \Sigma_Y$.*

This parallels the definition of continuous function in topology. Traditionally the phrase "measurable function" is used for a real-valued function but we will use it more generally. Our measurable functions have been sometimes called "measurable transformations". If we consider topological spaces with their Borel σ-field then any continuous function is clearly measurable. However many discontinuous functions are also measurable.

2.3 Metric Spaces and Convergence Properties of Measurable Functions

We now consider some special spaces and measurable functions to them. The most important measurable space is $\mathbf{R}$ with the Borel algebra generated by the usual topology. Unless we say otherwise we shall always mean "reals equipped with this topology and σ-field" when we say "reals". The following seemingly easy proposition gives some closure properties of real-valued measurable functions; it will turn out to conceal a subtlety that bears further scrutiny.

Proposition 2.12 *The absolute value of a measurable function is measurable. The sum and product of two measurable functions are measurable. The multiplicative inverse of an everywhere nonzero measurable function is measurable.*

Proof. The absolute value function is continuous and hence measurable, hence the first statement is proved. For the next statement the same argument should work since $+, * : \mathbf{R}^2 \longrightarrow \mathbf{R}$ are continuous. However, what this proves is that $+, *$ are measurable with respect to the product Borel algebra $\mathcal{B} \otimes \mathcal{B}$ on $\mathbf{R}^2$, which is not à priori the same as the Borel algebra generated by the product topology. In fact it is but the proof is not trivial and is the subject of the next proposition. To verify the last statement let $\mathbf{R}'$ be the reals with zero removed; let $\eta : \mathbf{R}' \longrightarrow \mathbf{R}$ be given by $\eta(x) = \frac{1}{x}$. Clearly η is continuous. Given f a nonzero measurable function we define $g(x) = \frac{1}{f(x)}$. By proposition 2.8 we need to verify that $g^{-1}(O)$ is measurable only for an open set O in $\mathbf{R}$. Now we have $g^{-1}(O) = g^{-1}(O \cap \mathbf{R}') = f^{-1}(\eta^{-1}(O \cap \mathbf{R}'))$; this is a measurable set since η is continuous and f is measurable. ∎

In order to fill the lacuna in the above proof we need a digression into metric spaces. Metric spaces play a central role in probability theory though one would not have guessed that from the initial definitions.

Definition 2.13 *A metric space with a countable, dense subset is called* **separable**.

Typical examples are $\mathbf{R}^n$. What separability does is give one access to a *countable* family of generating open sets

Proposition 2.14 *In a separable metric space, say (X, d), there is a countable family of open sets H_i such that every open subset may be written as the union of the H_i that it contains.*

Proof. Let $\{x_i\}$ be a dense subset of (X, d). Define $H_{i,m} = \{x | d(x, x_i) < 1/m\}$, where m is an integer. Let O be an open subset of X and let z be a point in O. Now there exists an integer n such that the open ball centred at z and with radius $1/n$ is contained in O. Since the set $\{x_i\}$ is dense there is a j such that $d(z, x_j) < 1/(2n)$. Then $H_{j,2n}$ is contained in O and includes the point z. ∎

Now we can relate the two Borel structures as required.

Proposition 2.15 *Let (X_1, d_1) and (X_2, d_2) be two separable, metric spaces. Let Y be the product equipped with the product topology. Then the Borel algebra of Y, written $\mathcal{B}_Y$, is the same as the product σ-field $\mathcal{B}_1 \otimes \mathcal{B}_2$.*

Proof. It is easy to see that Y is also separable; in fact we can construct the countable family of open sets by taking products of open balls. Thus we have a countable family of open rectangles which generate the topology of Y and hence the Borel algebra $\mathcal{B}_Y$. These are all in the σ-field $\mathcal{B}_1 \otimes \mathcal{B}_2$. Now any open set is the *countable* union of these special rectangular one sets and thus any open set in the product topology of Y is in $\mathcal{B}_1 \otimes \mathcal{B}_2$. Thus we have that $\mathcal{B}_Y \subseteq \mathcal{B}_1 \otimes \mathcal{B}_2$. Now consider the identity function on Y viewed as a function between measurable spaces, i.e. $i_Y : (Y, \mathcal{B}_Y) \longrightarrow (Y, \mathcal{B}_1 \otimes \mathcal{B}_2)$. This function is the one induced by universality of the product given the projections onto the components. Clearly the projections are continuous (by definition of the product topology) and hence measurable hence the map i_Y is measurable. This means that $\mathcal{B}_1 \otimes \mathcal{B}_2 \subseteq \mathcal{B}_Y$. ∎

We now discuss a theorem which shows the striking contrast between measure theory and topology. We take (X, Σ) to be a measurable space, (Y, d) to be a metric space with the induced Borel algebra $\mathcal{B}_Y$.

Definition 2.16 *Given a family of functions* $\{f_n : X \longrightarrow Y | n \in \mathbb{N}\}$ *we say that the family* **converges pointwise** *to* f *if* $\forall x \in X. \lim_{n\to\infty} f_n(x) = f(x)$.

Theorem 2.17 *If a family of measurable functions* $\{f_n : X \longrightarrow Y | n \in \mathbb{N}\}$ *converges pointwise to* f *then* f *is also measurable.*

Proof. We need to show that for an open set O of Y that $f^{-1}(O)$ is measurable. We will describe this set in terms of the f_n and use measurability of the f_n to establish measurability of f.

Given a subset O of Y and a point $y \in Y$ we define a distance function, also written d, by $d(y, O) = \inf_{u\in O} d(u, y)$. Let O be an open set in Y and let k be a positive integer, we define

$$O_k = \{u \in O | d(u, O^c) > 1/k\}.$$

Now it is not hard to see that the O_k are open sets[1], so we have an increasing sequence of open sets contained in O with $O = \cup_{k\in\mathbb{N}} O_k$. In fact we have $\overline{O_k} \subseteq O_{k+1}$.

Now fix a point x of X. Suppose that x is in $f^{-1}(O)$, then x has to be in $f^{-1}(O_k)$ for some k. We have $\lim_{n\to\infty} d(f_n(x), f(x)) = 0$ by hypothesis. Thus, for q large enough, say $\forall q \geq m$, we have $f_q(x) \in O_k$. Let $H^k_m = \cap_{q\geq m} f_q^{-1}(O_k)$. This set is measurable since all the f_n are and the countable intersection of measurable sets is measurable. Clearly we have $f^{-1}(O_k) \subseteq \cup_m H^k_m$. We define $G^k = \cup_m H^k_m$, which is also a measurable set being the countable union of measurable sets. Taking the union over all k gives $f^{-1}(O) = \cup_k f^{-1}(O_k) \subseteq \cup_k G^k \stackrel{\text{def}}{=} W$. Clearly W is measurable. Now we show that W is in fact $f^{-1}(O)$ to complete the argument. Let $u \in W$, i.e. there is an r such that $u \in G^r$, i.e. there is a j such that $u \in H^r_j$. Thus for all $q > j$ we have $u \in f_q^{-1}(O_r)$. Thus we have

$$f(u) = \lim_{q\to\infty} f_q(u) \in \overline{O_r} \subseteq O_{r+1} \subseteq O.$$

Thus $u \in f^{-1}(O)$ and $W = f^{-1}(O)$, thus f is measurable. ∎

This is a very typical argument and is worth ploughing through for that reason alone. The discussion can be turned around, one can show that every measurable function is the pointwise limit of very special functions. This fact lies at the heart of integration theory.

Definition 2.18 *A measurable function is called* **simple** *if its range is finite.*

For real valued functions we have the following very important result.

Theorem 2.19 *Given a non-negative measurable function* $f : X \longrightarrow \mathbf{R}$ *there is a family of simple functions* s_i *such that* $\forall i. s_i \leq s_{i+1} \leq f$ *and the* s_i *converge pointwise to* f.

[1] It is easy to verify that the complements are closed sets.

Proof. The strategy is to define step functions but on possibly complicated domains obtained by inverse images. For $n \in \mathbb{N}$ and $1 \leq i \leq n2^n$ define $E_{n,i} \stackrel{\text{def}}{=} f^{-1}([\frac{i-1}{2^n}, \frac{i}{2^n}])$ and $F_n \stackrel{\text{def}}{=} f^{-1}([n, \infty])$. Since f is measurable the sets E and F are also measurable. We set

$$s_n = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} \chi_{E_{n,i}} + n\chi_{F_n}.$$

It is easy to see that the s_n are increasing. Suppose that $f(x) = r$ for some x and suppose that $r \in [\frac{i-1}{2^n}, \frac{i}{2^n}]$. Then $x \in E_{n,i}$ and $s_n(x) = \frac{i-1}{2^n} \leq r$. Note that s_n stays within $\frac{1}{2^n}$ of $f(x)$ as n increases so s_n converges to f. If $f(x) = \infty$ for some x then $s_n(x) = n$ which will diverge as n goes to infinity. ∎

2.4 Measures

Measurable spaces or σ-fields are merely the arenas on which measure theory happens. The key notion of "measure" will now be introduced. Roughly speaking, a measure is an assignment of size to the sets in σ-field. This size is typically a real number but it could be a real number between 1 and 0, a probability measure, or an extended nonnegative real number, i.e. one from $[0, \infty]$, or even a complex number. These theories are all slightly different and play different roles in mathematics. For us the most important case will be probability measure but it is worth seeing what happens when ∞ is admitted as a possible value, this is of particular importance in integration theory.

Definition 2.20 *A* **measure** *(***probability measure***) μ on a measurable space (X, Σ) is a function from Σ (a set function) to $[0, \infty]$ ($[0, 1]$), such that if $\{A_i | i \in I\}$ is a countable family of pairwise disjoint sets then*

$$\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i).$$

In particular if I is empty we have

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

A set equipped with a σ-field and a measure defined on it is called a **measure space***.*

This property is called *countable additivity* or *σ-additivity*. It is possible to develop a theory with just finite additivity but many basic results are counterintuitive.

In the rest of this section we are always talking about a set X equipped with a σ-field Σ and a measure μ. We shall always mean "measurable set" when we just say "set". We use letters like A, B to stand for measurable sets.

Proposition 2.21 (Monotonicity and Continuity) *1. If $A \subseteq B$ then $\mu(A) \leq \mu(B)$.*

2. If $A_1 \subseteq A_2 \subseteq \ldots A_n \subseteq \ldots$ and $\cup_i A_i = A$ then $\lim_{i \to \infty} \mu(A_i) = \mu(A)$.

3. If $A_1 \supseteq A_2 \supseteq \ldots A_n \supseteq \ldots$ and $\cap_i A_i = A$ then $\lim_{i \to \infty} \mu(A_i) = \mu(A)$, if $\mu(A_1)$ is finite.

Proof. For (1) we note that $B = A \cup (B - A)$. Thus $\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A)$.

For (2) we define a family of pairwise disjoint sets B_i inductively as follows. We set $B_1 = A_1$ and $B_{i+1} = A_{i+1} - A_i$. Since we have $\cup_i B_i = A$ and the B_i are pairwise disjoint, we have, by countable additivity, that $\mu(A) = \sum_i B_i$. But we also have $A_n = \bigcup_{i=1}^{n} B_i$, thus $\mu(A_n) = \sum_{i=1}^{n} \mu(B_i)$. Thus

$$\lim_{n \longrightarrow \infty} \mu(A_n) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} \mu(B_i) = \sum_{i=1}^{\infty} \mu(B_i) = \mu(A).$$

A similar argument holds for the third clause though the caveat about $\mu(A_1)$ being finite is essential. ∎

The following corollary is immediate.

Corollary 2.22 (Convexity) *For any countable family of sets B_i we have $\mu(\cup_i B_i) \leq \sum_i \mu(B_i)$.*

The first example looks natural but is pathological and is an important source of counterexamples.

Example 2.23 For X an infinite set we define a measure on the powerset of X by setting $\mu(A)$ equal to the number of elements of A if A is finite and ∞ otherwise. This measure is called *counting measure*. Many small variations are possible, such as weighting the points of X differently.

The next example appears artificial but is of central importance.

Example 2.24 Fix a set X and a point x of X. We define a measure, in fact a probability measure, on the σ-field of all subsets of X as follows. We use the slightly peculiar notation $\delta(x, A)$ to emphasize that x is a parameter in the definition.

$$\delta(x, A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

This measure is called the *Dirac delta measure*. Note that we can fix the set A and view this as the definition of a (measurable) function on X. What we get is the characteristic function of the set A, χ_A.

In order to cope with ∞ as a legitimate value in measure theory we need to adopt some arithmetic conventions for dealing with it. We decree that $0 \cdot \infty = 0$, otherwise we adopt the reasonable rules, $a \cdot \infty = \infty \cdot a = \infty$ if $a \neq 0$ and $a + \infty = \infty + a = \infty$. This choice gives commutativity and associativity of addition and multiplication as well as unrestricted distributivity. Furthermore if we have two sequences $a_n \longrightarrow a$ and $b_n \longrightarrow b$ we will have $a_n \cdot b_n \longrightarrow a \cdot b$ whatever a or b might be.

A measure space (X, Σ, μ) is said to be **finite** if $\mu(X) < \infty$ and σ**-finite** if X can be written as a countable union of sets if finite measure. Finite measure spaces arise in probability theory but many results actually hold for the σ-finite case and in these notes we will not always assume finiteness.

Proving the Strong Law We now have the technical tools to prove the strong law. First of all recall that the sequence space is Ω. The σ-field on Ω is defined as follows. We consider all sets of form

$$A_s := \{s \cdot s' : s \text{ a finite sequence, } s' \in \Omega\}$$

where $\cdot$ stands for concatenation of sequences. We consider the σ-field generated by these sets. We assign a probability measure Pr as follows. For any set of the form A_s we associate the probability $2^{-\text{length}(s)}$.

Recall that we defined the set $A \subseteq \Omega$ by

$$A := \{s : \lim_{n \longrightarrow \infty} [\frac{1}{n} \sum_{i=1}^{n} s_i] = \frac{1}{2}\}.$$

The strong law says that the probability of landing in the set A is 1 if we choose an infinite sequence randomly. It will be more convenient to prove that $E := A^c$ has probability 0.

ProofOf the Strong Law. We define the measurable functions $X_j : \Omega \longrightarrow \mathbf{R}$ by

$$X_j(s) = \begin{cases} 1 & \text{if } s_j = H \\ 0 & \text{if } s_j = T. \end{cases}$$

We define the function $S_n(s) := \sum_{j=1}^{n} X_j(s)$. Then we can define A as the set $\{s : \lim_{n \longrightarrow \infty} \frac{S_n(s)}{n} = \frac{1}{2}\}$.

We claim that $\lim_{n \longrightarrow \infty} \frac{S_n(s)}{n} = \frac{1}{2}$ if and only if $\lim_{m \longrightarrow \infty} \frac{S_{m^2}(s)}{m^2} = \frac{1}{2}$. To see this, for any n we choose m by $m^2 \leq n \leq (m+1)^2$, or, equivalently, $0 \leq n - m^2 \leq 2m$. If we choose m this way we have

$$|\frac{S_n}{n} - \frac{S_{m^2}}{m^2}| = |\frac{S_n}{m^2} - \frac{S_{m^2}}{m^2} + (\frac{1}{n} - \frac{1}{m^2})S_n|$$

$$\leq \frac{|n - m^2|}{m^2} + n|\frac{1}{n} - \frac{1}{m^2}| \leq \frac{2}{m} + \frac{2}{m} = \frac{4}{m}.$$

Thus as $n, m \longrightarrow \infty$ the difference between $\frac{S_n}{n}$ and $\frac{S_{m^2}}{m^2}$ goes to 0. Hence we can now work with $\frac{S_{m^2}}{m^2}$ in reasoning about E.

The strategy of the proof is to consider sets of sequences, E_ϵ, where the asymptotic behaviour is bounded away from 0 by ϵ. The set E is a countable union of such sets. We will show that each of the E_ϵ has probability 0, thus E will also have probability 0. To show that each E_ϵ has probability 0 we will use the limiting properties of the probability.

We define $E_\epsilon := \{s : \overline{\lim} |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}$. We define

$$E_a^b := \bigcup_{m=a}^{b} \{s : |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}.$$

The set E_a^b represents the sequences where the inequality $|\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon$ is satisfied at least once for m in between a and b. Finally we define

$$E_a = \bigcup_{b=a}^{\infty} E_a^b.$$

By the basic convexity property we have

$$Pr(E_a^b) \leq \sum_{m=a}^{b} Pr(\{s : |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}).$$

Now using Chebyshev's inequality just as we did in the proof of the weak law we get

$$Pr(E_a^b) \leq \frac{1}{4\epsilon^2} \sum_{m=a}^{b} \frac{1}{m^2}.$$

As we let b go to infinity and use the fact that for fixed a the family E_a^b is a nested increasing family of sets so that by continuity we get

$$Pr(E_a) = \lim_{b \longrightarrow \infty} Pr(E_a^b) \leq \frac{1}{4\epsilon^2} \sum_{m=a}^{\infty} \frac{1}{m^2}.$$

What does it mean for a sequence to be in a set of the form E_a? It means that at some stage *after* a the inequality $|\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon$ is satisfied. Now, by the definition of $\overline{\lim}$ to be in E_ϵ this inequality must be satisfied infinitely often, i.e. we have $E_\epsilon = \cap_a E_a$. As a inceases we get a nested decreasing family of sets so by continuity again we have $Pr(E_\epsilon) = \lim_{a \longrightarrow \infty} Pr(E_a)$. Thus

$$Pr(E_\epsilon) \leq \lim_{a \longrightarrow \infty} [\frac{1}{4\epsilon^2} \sum_{m=a}^{\infty} \frac{1}{m^2}] = 0.$$

The last equality follows from the fact that we are looking at the tails of a convergent sequence. Thus we have established that for any $\epsilon > 0$, $Pr(E_\epsilon) = 0$.

Now consider the set E. This set can be written as the countable union

$$E = \bigcup_{k=1}^{\infty} E_{\frac{1}{k}}$$

where k is an integer. Thus since each $E_{\frac{1}{k}}$ has probability 0 we get that E has probability 0, or A has probability 1. ∎

A set of measure 0 is sometimes called a **negligible** set. They play a very important role in the theory and one often hears phrases like "almost everywhere" or "almost surely". What they usually mean is that something or other is true except on a negligible set. To be sure the notion of negligible depends on the measure so one should be careful in interpreting such phrases. One very annoying feature of negligible sets is that they may contain nonmeasurable subsets, whereas we would certainly like to say that all the subsets of a negligible set are negligible as well. This would be true if we could be certain that all the subsets of a negligible set were measurable. Fortunately this can be fixed by "completing the measure". One cannot just throw in all subsets of the negligible sets but what turns out to work is to throw in all sets that are sandwiched between two sets of the same measure. We proceed to make all this precise.

Definition 2.25 *A measure space (X, Σ, μ) is said to be* **complete** *if every subset of a negligible set is in Σ.*

Theorem 2.26 *Given (X, Σ, μ), there is a σ-field, $\Sigma' \supseteq \Sigma$ and a measure μ' on Σ' such that (X, Σ', μ') is complete and $\mu(A) = \mu'(A)$ for any A in Σ.*

Proof. We define the extended σ-field as follows

$$\Sigma' = \{Z \in \mathcal{P}(X) | \exists A, B \in \Sigma . A \subseteq Z \subseteq B \wedge \mu(B - A) = 0\}.$$

Clearly $\Sigma' \supseteq \Sigma$ but we need to verify that it is indeed a σ-field. Let Z be in Σ'. Then we have A, B as above. Since $A \subseteq Z \subseteq B$ we have $B^c \subseteq Z^c \subseteq A^c$ and, of course, A^c and B^c are both in Σ. Also $\mu(A^c - B^c) = \mu(B - A) = 0$ so Z^c is in Σ'. Now suppose that Z_n is a countable family of sets in Σ'. We have two countable families of sets A_n and B_n both drawn from Σ with

$$\forall n. A_n \subseteq Z_n \subseteq B_n, \mu(B_n - A_n) = 0.$$

We set $Z = \cup_n Z_n$, $A = \cup_n A_n$ and $B = \cup_n B_n$. Clearly A and B are in Σ and $A \subseteq Z \subseteq B$. Now $(B - A) \subseteq \cup_n [B_n - A_n]$ hence by monotonicity of measure and convexity we have

$$\mu(B - A) \leq \mu(\cup_n [B_n - A_n]) \leq \sum_n \mu(B_n - A_n) = 0.$$

Thus Z is in Σ' and we have verified that Σ' is a σ-field.

We define μ' as follows. First, note that $A \subseteq B$ and $\mu(B - A) = 0$ implies that $\mu(A) = \mu(B)$. Given $A \subseteq Z \subseteq B$ we define $\mu'(Z) = \mu(A) = \mu(B)$. We need to verify that this is independent of the chosen A and B. Suppose that we have another pair $A' \subseteq Z \subseteq B'$ with $\mu(B' - A') = 0$. Now since $B' \supseteq Z \supseteq A$ and $B \supseteq Z \supseteq A'$ we have $\mu(B') \geq \mu(A) = \mu(B)$ and $\mu(B) \geq \mu(A') = \mu(B')$. Thus we have $\mu(A) = \mu(A') = \mu(B) = \mu(B')$ and μ' is well defined. Finally we need to verify countable additivity of μ'. Suppose that we have a pairwise disjoint family of sets Z_n from Σ'. Then the associated sets A_n are also pairwise disjoint thus we have

$$\mu'(\cup_n Z_n) = \mu(\cup_n A_n) = \sum_n \mu(A_n) = \sum_n \mu'(Z_n).$$

To verify that μ' is complete we proceed as follows. Suppose that $Z \in \Sigma'$, $Z' \subseteq Z$ and that we have $\mu'(Z) = 0$. Then we have sets $A, B \in \Sigma$ with $A \subseteq Z \subseteq B$ and with $\mu(B) = 0$. Now we have $\emptyset \subseteq Z' \subseteq B$ thus Z' is in Σ' and has μ'-measure 0. ∎

Remark 2.27 The word "completion" is not a very good choice. It suggests that the completion is unique[2] but it is not in fact. It is possible to give simple (finite) examples where the completion process can be further extended. What is true is that the completion process described in the proof above, if carried out a second time, would yield no new sets.

Given a measure space (X, Σ, μ) one often implicitly talks about the completion with respect to μ rather than the given σ-field. It is conventional when talking about the reals to use the phrase "Borel field" or "Borel sets" to refer to the σ-field generated by the open intervals and the phrase "Lebesgue measurable sets" to talk about the sets that arise from the completion process above with respect to the standard Lebesgue measure, defined below.

2.5 Lebesgue Measure

We now proceed to construct the most important single example of a measure, the Lebesgue measure on the real line. We want a notion of measure on the real line to assign the familiar notion of length to intervals. This simple requirement immediately brings us face-to-face with the technical subtleties that lie at the foundations of set theory.

[2]Indeed some reputable books say that it is.

Theorem 2.28 (Vitali) *Assuming the axiom of choice, there is no measure on the real numbers which is translation-invariant, assigns a measure to all sets and which assigns a nonzero value to the interval* $[0, 1]$.

We defer the proof to the end of the section in order to not interrupt the flow of ideas.

Let us start naively. We would like a measure which agrees with the usual notion of length of an interval. Recall that $[0, \infty]$ is a complete lattice so we can take lubs and glbs as we please. Let us call our putative measure m and we begin by trying to define it on all subsets of $\mathbf{R}$. We begin as follows

$$m((a, b)) = m((a, b]) = m([a, b)) = m([a, b]) = b - a$$

where a and b are real numbers. If we have a pairwise disjoint collection of intervals $\{I_k | k \in \mathcal{K}\}$ we define

$$m(\bigcup_{k \in \mathcal{K}} I_k) = \sum_{k \in \mathcal{K}} m(I_k)$$

as suggested by the countable additivity requirement. It is not hard to check that m does not depend on how we partition a set into disjoint intervals. Now suppose that we have an arbitrary subset A of $\mathbf{R}$. We define a putative measure μ^* as follows. Let S be a family of intervals that covers A. We set $\mu^*(A)$ to be the infimum over all families S that cover A, of $m(S)$. This is indeed defined on all subsets of $\mathbf{R}$. Unfortunately it is not a measure because it is not countably additive. Nevertheless it is an interesting set function which will be the basis of our construction of Lebesgue measure.

As a prelude let us calculate μ^* for some sets. Clearly $\mu^*([a, b]) = b - a$ and $\mu^*(\mathbf{R}) = \infty$. We claim that $\mu^*(\mathbb{Q}) = 0$ where $\mathbb{Q}$ is the set of rational numbers. Let ϵ be any positive number, we will cover the rationals with a family of intervals of total length ϵ. Let $r_1, r_2, \ldots$ be any enumeration of the rationals. Consider the family of intervals

$$\{[r_1 - \frac{\epsilon}{4}, r_1 + \frac{\epsilon}{4}], [r_2 - \frac{\epsilon}{8}, r_2 + \frac{\epsilon}{8}], \ldots, [r_j - \frac{\epsilon}{2^{j+1}}, r_j + \frac{\epsilon}{2^{j+1}}], \ldots\}$$

these cover the rationals and have total length ϵ. Thus the inf is clearly 0. This calculation depends very much on the rational being countable and shows that μ^* is 0 is zero for any countable set. However there are uncountable sets which get assigned 0 by μ^*; the reader is invited to check this for the Cantor set.

What properties does μ^* have? It is clearly monotone but not countably additive. We claim that it satisfies the convexity property

$$\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$$

where $\{A_i\}$ is a countable family not necessarily pairwise disjoint. To see this we adopt a common trick in analysis. We want to prove that some quantity, say K, is less than some other quantity, say L, both defined through a limiting process. It is often easier to prove the equivalent statement that for any ϵ, no matter how small, that $K \leq L + \epsilon$; the ϵ gives one room to manoeuvre. In detail we proceed as follows. Suppose that one does have an arbitrary positive ϵ. Since $\mu*$ is defined as the infimum of m over all covers we can find a family of intervals $\mathcal{I}_1$ covering A_1 with $m(\mathcal{I}_1) \leq \mu^*(A_1) + \frac{\epsilon}{2}$. We have written $m(\mathcal{I}_1)$ as a shorthand for the sum of the lengths of all the intervals in the family $\mathcal{I}_1$. Similarly we can find a family $\mathcal{I}_n$ covering A_n with $m(\mathcal{I}_n) \leq \mu^*(A_n) + \frac{\epsilon}{2^n}$. Notice how we have used the ϵ; normally we would use the definition of μ^* as an inf to conclude that $\mu^*(A_1) \leq m(\mathcal{I}_1)$ (which is, of course, true but useless) but the extra "room" given by the epsilon allows us to get an inequality in the reverse direction. Adding up all the inequalities we get $m(\mathcal{I}) \leq \sum_i \mu^*(A_i) + \epsilon$ where the family $\mathcal{I}$ is the union of all the families $\mathcal{I}_n$. Clearly $\mathcal{I}$ covers $\cup_i A_i$ so $\mu^*(\cup_i A_i) \leq m(\mathcal{I}) \leq \sum \mu^*(A_i) + \epsilon$. Since this holds for any ϵ we have $\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$.

We abstract out the properties of μ^* in a definition.

Definition 2.29 *An* **outer measure** *on a set X is a set function $\mu^* : \mathcal{P}(X) \longrightarrow [0, \infty]$, defined on all subsets of X, such that*

1. $\mu^*(\emptyset) = 0$,
2. $A \subseteq B \Rightarrow \mu^*(A) \leq \mu^*(B)$ *and*
3. *for any countable family of subsets of X, say $\{A_i\}$, we have*

$$\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i).$$

We have shown that the example μ^* we have constructed on the reals is an outer measure. This is still not a measure but it is defined on all sets. It turns out that there is a general construction which produces a σ-field and a measure defined on it from an outer measure; the σ-field will not, in general, be the collection of all sets. We have to choose those sets that behave nicely with respect to the outer measure when we take complements. Roughly speaking the σ-field consists of those sets which can be used to "split" any other sets. Applied to our example it produces Lebesgue measure.

Theorem 2.30 *Let X be a set and let μ^* be an outer measure defined on X. Denote by Σ the collection of all subsets, say A, such that for every subset E of X we have*

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

For all A in Σ define $\mu(A) = \mu^*(A)$. Then (X, Σ, μ) is a measure space.

Proof. It is trivial to verify that $\emptyset$ is in Σ and that Σ is closed under complementation. In order to verify countable additivity we first verify finite additivity.

Suppose that $A, B \in \Sigma$. We have, for any subset E of X,

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

Now we note that

$A \cap B^c = (A \cap B)^c \cap A$, hence
$(A \cap B^c) \cap E = ((A \cap B)^c \cap A) \cap E$, hence
$(A \cap B^c) \cap E = ((A \cap B)^c \cap E) \cap A$.

Similarly we have

$A^c \cap E = ((A \cap B)^c \cap A^c) \cap E = ((A \cap B)^c \cap E) \cap A^c.$

Using these set-theoretic identities to rewrite the last two terms in the expression for $\mu^*(E)$ we get

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

Using the fact that A is in Σ we can combine the last two terms

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

The last equality shows that $A \cap B$ is in Σ. Now an easy induction shows that Σ is closed under finite intersections (and hence unions too).

We next show that Σ is closed under countable unions using essentially a limiting version of the finite additivity argument. Note, however, that it suffices to prove this for pairwise disjoint families. In this case we can use the following simple identity. If A, B are disjoint sets in Σ and E is any set then

$$\mu^*((A \cup B) \cap E) = \mu^*(A \cap E) + \mu^*(B \cap E).$$

This is an easy exercise and trivially extends to any finite pairwise disjoint family. Suppose that we have a countable pairwise disjoint family $\{A_i | i \in \mathbb{N}\}$ in Σ. The union of any finite subfamily, in particular $\{A_i | i \leq n\}$, is in Σ.

Using the defining property of Σ for any set E we have

$\mu^*(E) = \mu^*((\cup_{i=1}^n A_i) \cap E) + \mu^*((\cup_{i=1}^n A_i)^c \cap E)$

repeatedly using the simple identity of the last paragraph we get

$= \sum_{i=1}^n \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^n A_i)^c \cap E).$

Now we use monotonicity of μ^* to get

$\mu^*(E) \geq \sum_{i=1}^n \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E).$

Since this is true for any $n \in \mathbb{N}$ we have

$\mu^*(E) \geq \sum_{i=1}^\infty \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E).$

Now using subadditivity of outer measures we get

$\geq \mu^*((\cup_{i=1}^\infty A_i) \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$

and subadditivity again gives

$\geq \mu^*(E).$

All these inequalities must therefore be equalities and we have

$$\mu^*(E) = \mu^*((\cup_{i=1}^\infty A_i) \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$$

which establishes that $\cup_i A_i$ is in Σ. Thus Σ is indeed a σ-field.

We need to show that μ, i.e. μ^* restricted to the sets in Σ, is countably additive, i.e. is a measure. In the preceding calculation we have a countable pairwise disjoint family $\{A_i\}$ of sets in Σ. We take E to be $\cup_i A_i$ in the equation

$$\mu^*(E) = \sum_{i=1}^{\infty} \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$$

to get

$$\mu(\cup_i A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

where the second term drops out since it is just μ^* of the empty set. ∎

The theorem just proved allows one to cut down an outer measure defined on all sets to a measure defined on a σ-field. Another very useful type of theorem constructs a measure on a σ-field by starting with data on a restricted family of sets that generate the σ-field. We state a typical theorem of this kind.

Definition 2.31 *A family $\mathcal{F}$ of subsets of X is called a* **semi-ring** *if*

1. $\emptyset \in \mathcal{F}$,
2. $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$ *and,*
3. *if $A \subseteq B$ are in $\mathcal{F}$ then there are* finitely *many* pairwise disjoint *subsets $C_1, \ldots, C_k \in \mathcal{F}$ such that $B - A = \cup_{i=1}^{k} C_i$.*

This is not the form of the definition that one is used to in algebra because of the strange last condition but this is precisely the property that holds for "hyperrectangles" in $\mathbf{R}^n$.

Theorem 2.32 *Suppose that $\mathcal{F}$ is a semi-ring on X and $\mu : \mathcal{F} \longrightarrow [0, \infty]$ satisfies*

1. $\mu(\emptyset) = 0$,
2. *μ is finitely additive and*
3. *μ is countably subadditive.*

Then μ extends to a measure on the σ-field generated by $\mathcal{F}$.

Countably subadditive means that $\mu(\cup_n A_n) \leq \sum_n \mu)A_n)$. The proof of this theorem may be found in a standard text on probability and measure, for example the book by Kingman and Taylor [KT66] or the one by Billingsley [Bil95] or the one by Ash [Ash72]. The intervals on the reals form a typical example of a semi-ring and the length function satisfies the conditions of the theorem so this extension theorem gives another construction of Lebesgue measure.

2.6 Nonmeasurable sets

It is easy to give artificial examples of nonmeasurable sets. In this section we give a construction due to Vitali which produces a nonmeasurable subset of the reals.

Theorem 2.33 *There are Lebesgue nonmeasurable subsets of the real line.*

The proof depends on the translation invariance of Lebesgue measure. In other words if we have a measurable set E and we define $E + a := \{x + a \mid x \in A\}$ then $\mu(E + a) = \mu(E)$. In fact the proof below shows that there is no translation-invariant measure on the reals except for trivial examples (such as the one which assigns zero measure to all sets).

Proof. Let $\mathbf{R}$ stand for the real line. We define an equivalence relation on $\mathbf{R}$ as follows. We say $x \sim y$ for $x, y \in \mathbf{R}$ if $x - y$ is rational. Note that for every x in $\mathbf{R}$ there is a $y \in (0,1)$ with $x \sim y$. Let E be the set obtained by choosing from every $\sim$ equivalence class one element in $(0,1)$. We claim that such an E cannot be measurable.

Now let $E + r$ stand for the set $\{x + r : x \in E\}$ where r is a rational number. Now if r, r' are distinct rational numbers the sets $E + r$ and $E + r'$ are disjoint. Furthermore for any x in $(0,1)$ there is some rational number $r \in (-1,1)$ such that $x \in E + r$. To see this note that for any x in $(0,1)$ we have an element of E, say u, with $x \sim u$. Now let $r = x - u$, which must be a rational number. Then $x \in E + r$. Since both x and u are in $(0,1)$ we must have $r \in (-1,1)$.

Now consider the countable union

$$S = \bigcup_{r \in (-1,1)} E + r$$

where the r are all rational. The set S contains the open interval $(0,1)$ and is contained in the open interval $(-1,2)$. Furthermore all the $E+r$ are translates of E. Thus if μ is any translation-invariant measure on $\mathbf{R}$ we have $\mu(E) = \mu(E+r)$ for any r. Suppose that $\mu(E) = \alpha$ then $\mu(S) = \Sigma_{r\in(-1,1)}\alpha$ which is either 0 (if $\mu(E)$ is 0) or ∞. Now it is clear that

$$\mu((0,1)) \leq \mu(S) \leq \mu((-1,2)).$$

If in addition we want μ to agree with the lengths of intervals we have $1 \leq \mu(S) \leq 3$. But this contradicts $\mu(S) = 0$ or ∞. ∎

The manifest use of the axiom of choice raises the question of whether there are models of ZF which do not validate the axiom of choice and for which all subsets of the reals are measurable. One can drop the axiom of choice and construct models of ZF set theory in which the axiom of choice is false and all subsets of the reals are Lebesgue measurable. On the other hand the resulting universe has some rather unpleasant properties. For example the reals can be expressed as a countable union of countable sets. If we want a healthier model one can try to construct models satisfying the principle of dependent choice (DC) which is a little weaker than the axiom of choice. Solovay [Sol70] has indeed constructed a model of ZF set theory with DC such that all subsets of the reals are measurable. His construction is based on the assumption that there exists an inaccessible cardinal. The real point is that we are not likely to encounter nonmeasurable sets in the course of normal mathematical activity, but one has to be aware that they exist. A good discussion of these matters is in the book "Lectures in Set Theory" by Jech [Jec71]

2.7 Exercises

1. Prove proposition 2.2.

2. Prove proposition 2.7.

3. Prove proposition 2.8.

4. Is there an infinite σ-field with only countably many members?

5. Suppose that (X, Σ, μ) is a measure space and $A_1, A_2 \in \Sigma$. Prove that $\mu(A_1 \cup A_2) \leq \mu(A_1) + \mu(A_2)$.

6. Suppose that $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq \ldots$ are all measurable subsets of some measure space. Prove
$$\lim_{n\to\infty} \mu(A_n) = \mu(\cup_i A_i).$$
use this to prove a countable version of the convexity result of part 2.

7. Give an example (finite please!) of a set and σ-field and a measure such that the completion with respect to the measure can be consistently extended to yet more sets.

8. Prove that the Cantor set has Lebesgue measure zero.

9. Prove that there are only $\aleph_1$ Borel sets but $\aleph_2$ Lebesgue measurable sets.

10. Prove the "simple formula" used in the proof of theorem 2.30. If A, B are disjoint sets in Σ and E is any set then
$$\mu^*((A \cup B) \cap E) = \mu^*(A \cap E) + \mu^*(B \cap E).$$

11. Show that the measure space produced by theorem 2.30 is complete.

12. Suppose that X is an uncountable set and μ^* is an outer measure on X. We define μ^* by
$$\mu^*(E) = \begin{cases} 0 & \text{if } E \text{ is countable} \\ 1 & \text{otherwise.} \end{cases}$$
Describe the measurable sets and the measure.

Chapter 3

Integration

In this chapter we study Lebesgue integration and its relationship to the more familiar Riemann integration. In fact the two are intimately related. Before embarking on the formal study of integration it is worth recalling Lebesgue's own imagery for the difference between the two integrals; the following is taken from a lecture he gave on the origins of the Lebesgue integral.

Imagine a shopkeeper who keeps track of the money he receives in the course of a day. He could keep track of it by maintaining a running total as he receives the money or he could keep track of it by tossing the various coins and notes in bins and add up the totals later. The latter allows him to deal with money coming in much more quickly and is not affected by the money coming in an irregular fashion. The former method corresponds to Riemann integration while the latter to Lebesgue integration.

In Riemann integration we approximate the integral, say $\int_a^b f(x)dx$ by sums of the form $\sum_i f(x_i)(x_i - x_{i-1})$. We divide up the domain; the interval $[a, b]$ is partitioned by the x_i. Now this notion of integration is intuitive, has a strong constructive flavour and is the basis of all the approximation schemes that one encounters in numerical analysis. The main defect is that the function f cannot vary "too wildly". In Lesbegue integration one partitions the *range*. Now the functions can vary much more wildly but the sets involved can be much more complicated than intervals; they can be arbitrary measurable sets.

It might seem that the whole point of Lebesgue is to integrate pathological functions. How likely are these functions to arise in practice? Actually the real reason for the predominance of the Lebesgue integral in modern analysis is that it has a much smoother theory. In particular, under quite general conditions one can interchange limits and integration; i.e. in some appropriate sense Lebesgue integration is "limit preserving" but Riemann integration is not.

In this chapter we will approach the Lebesgue integral fairly abstractly, i.e. we will not need details of Lebesgue measure in order to define it. We will also discuss the relationship between the Lebesgue integral and the Riemann integral.

3.1 The Definition of Integration

How do we divide up the range? Recall the *simple functions* which we introduced at the end of the last chapter. These are measurable functions whose range is a finite set; thus they serve to define a finite partition of the range space. They will be the functions that we start our definition with and from there we extend to the general measurable functions by a limiting process. Recall that all measurable functions are limits of sequences of simple functions.

Suppose that we have a simple function from the reals to the reals, say s, whose range is the set $\{a_1, \ldots, a_n\}$. We define $\forall i \in \{1, \ldots, n\}. A_i \stackrel{\text{def}}{=} s^{-1}(a_i)$. The A_is are measurable sets if s is a measurable function but they could be quite complicated otherwise; far more complicated than the intervals that arise in Riemann integration. The natural definition for the integral of s is

$$\int s \, \mathrm{d}\mu = \sum_i a_i \mu(A_i)$$

where μ is Lebesgue measure. This pleasant picture is complicated by questions of well definedness immediately. What if $\mu(A_1)$ is infinite and $a_1 > 0$? It is reasonable to assign the value ∞ to the integral in this case, but what if, in addition, $\mu(A_2) = \infty$ and $a_2 < 0$? Thus it is entirely possible for a function to be measurable but not have a sensible integral.

Definition 3.1 *We say that a simple function s is* **integrable** *if whenever a is in the range of s, $a \neq 0 \Rightarrow \mu(s^{-1}(a)) < \infty$.*

It is possible to define integrable to mean that the sums arising in the definition of the integral are well-defined but the present definition is the usual one in the analysis literature. We can now make the proposed definition above official.

Definition 3.2 *Suppose that (X, Σ, μ) is a measure space and that $s : X \longrightarrow \mathbf{R}$ is an integrable, simple function with range $\{a_1 \ldots, a_n\}$. We say that the* **integral** *of s over X with respect to the measure μ is $\int_X s\mu = \sum_{i=1}^n a_i \mu(s^{-1}(a_i))$.*

Before we move on to the definition of the integral of more general functions we need to observe some possible pathologies that have to be dealt with. Consider, for definiteness, functions from $\mathbf{R}$ to $\mathbf{R}$. We might try to make use of the fact that every measurable function is the pointwise limit of a sequence of simple functions to define the integral of general measurable functions. Now suppose that we have the family $s_i(x) = 1$ if x is between i and $i + 1$ and 0 otherwise. The limit of this family is the everywhere 0 function whose integral should certainly be 0. However the integral of every function in the family is 1. It is important therefore that we use approximation from below. Thus if we have an increasing family of simple functions $s_1 \leq \ldots \leq s_i \leq s_{i+1} \leq \ldots$ with f as the supremum we can hope to define the integral of f as the limit of the integrals of the s_is.

Now consider the function $\frac{-1}{1+x^2}$. We would like to say

$$\int_{-\infty}^{\infty} \frac{-1}{1 + x^2} \, \mathrm{d}x = -\pi$$

since, as we no doubt recall from high-school, that

$$\int_{-\infty}^{\infty} \frac{1}{1 + x^2} \, \mathrm{d}x = \pi.$$

Unfortunately if we use our notion of approximation by simple functions *from below* we have problems since clearly a simple function below this function is going to be everywhere nonzero and hence not integrable. Thus it is clear that we should try to approximate negative functions from above and positive functions from below. A general function has to be split into its positive and negative parts and have the integrals defined separately for each part.

For the rest of this section we fix a measure space (X, Σ, μ). When we say "real-valued function" we will mean measurable real-valued function defined on X. Suppose f is a real-valued function defined on X. We write $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$; clearly both f_+ and f_- are measurable if f is.

Definition 3.3 *Suppose that f is an everywhere nonnegative real-valued function. We say that f is* **integrable** *if the everywhere nonnegative simple functions less than f are integrable and their integrals are bounded. If f is integrable we define*

$$\int_X f\mu = \sqcup \int_X s\mu$$

where the sup is over all nonnegative simple functions below f. If we have a measurable function g which takes on both positive and negative values we say that g is integrable if both g_+ and g_- are integrable and we set

$$\int_X g\mu = \int_X g_+\mu - \int_X g_-\mu.$$

Example 3.4 We take as our measure space (X, Σ, δ_x) where δ_x is the Dirac measure concentrated at the point x of X. Let f be any nonnegative real-valued function. We claim that

$$\int_X f \mathrm{d}\delta_x = f(x).$$

Note that the simple function $s(x) = f(x)$ and 0 everywhere else is a simple function below f. The integral of s wrt δ_x is $f(x)$. Furthermore any simple function t below f has the integral $t(x) \leq f(x)$. Thus the sup of the integrals of all the simple functions below f is precisely $f(x)$.

The next example is the standard advertisement for the superior generality of the Lebesgue integral.

Example 3.5 Let $f : [0, 1] \longrightarrow \mathbf{R}$ be given by $f(x) = 0$ if x is rational and $f(x) = 1$ if x is irrational. This f is in fact a simple function, in fact it is even the characteristic function of a measurable set. Thus its integral is just the measure of the irrationals between 0 and 1 which is 1.

One has to be careful about how the word "integrable" is used. Its use suggests that a function that is not integrable cannot have a sensible integral assigned to it. The definition is rather conservative and often people would like to say that certain integrals are defined but "divergent". Thus, for example, it is common to say that $\int_{-\infty}^{\infty} 1dx = \infty$. One uses the phrase "has a definite integral" for the less stringent. Thus one says that the function $\lambda x.1$ is not integrable but has definite integral between $-\infty$ and ∞ of ∞.

Example 3.6 The identity function on the reals is measurable (even continuous) but not integrable. This is a more troublesome example. A so-called pragmatic view of this is that the integral is 0; for example one can argue by symmetry. While this is what physicists and engineers usually say the correct statement is that this function is not integrable and does not have a definite integral over the given range.

3.2 Properties of the Integral

Now we can prove some basic properties of the integral. It is customary to introduce the notation $\int_A f\mu$ for the integral of f restricted to the measurable subset A of X with the induced measure. In the next proposition functions and integrals are always on X and f, g are used for integrable functions.

Proposition 3.7 *1. If $0 \leq f \leq g$ then $\int f\mu \leq \int g\mu$.*

2. If $0 \leq f$ and $0 \leq c$ is a constant then $\int cf\mu = c\int f\mu$.

3. $\int_A f\mu = \int_X f\chi_A\mu$ where χ_A is the characteristic function of A.

The proofs are elementary and are left to the exercises. The last is a triviality. The next proposition is the first step towards establishing linearity of the integral.

Proposition 3.8 *Let s, t be simple integrable functions. Then*

$$\int (s+t)\mu = \int s\mu + \int t\mu.$$

Proof. Let A be a measurable set, s any simple integrable function and define $\phi(A) = \int_A s\mu$. Then ϕ is a measure. To see this recall that any simple function can be written in the form of a finite sum

$$s = \sum_{i=1}^{n} a_i \chi_{A_i}$$

where for definiteness we assume that the a_i are distinct. Now suppose that $C = \cup_{i\in\mathbf{N}} C_i$ where the C_i are pairwise disjoint and measurable. We have

$$\begin{aligned}
\phi(C) &= \textstyle\sum_{i=1}^{n} a_i\mu(A_i \cap C)\\
&= \textstyle\sum_{i=1}^{n} a_i \sum_{j=1}^{\infty} \mu(A_i \cap C_j)\\
&= \textstyle\sum_{j=1}^{\infty} \sum_{i=1}^{n} a_i\mu(A_i \cap C_j)\\
&= \textstyle\sum_{j=1}^{\infty} \phi(C_j)
\end{aligned}$$

where the first equality is by definition of ϕ, the second by countable additivity, the third by absolute convergence and the fourth by definition. Thus we have shown that ϕ is countably additive.

Let t be written as $t = \sum_{j=1}^{m} b_j\chi_{B_j}$ where the b_j are all distinct. We set A_0 to be the set of all points where s vanishes and B_0 to be the set of all points where the t vanishes. Let $C_{ij} = A_i \cap B_j$. The sets C_{ij} form a partition of X, if we let the index values include 0. Now $s + t$ is simple and integrable. Thus $\psi(D) = \int_D (s+t)\mu$ is a measure. Thus we can write

$$\begin{aligned}
\textstyle\int_X (s+t)\mu &= \textstyle\sum_{i,j} \int_{C_{ij}} (s+t)\mu\\
&= \textstyle\sum_{i,j} (a_i + b_j)\mu(C_{ij})\\
&\textstyle\sum_{i,j} (\int_{C_{ij}} s\mu + \int_{C_{ij}} t\mu)\\
&= \textstyle(\sum_{i,j} \int_{C_{ij}} s\mu) + (\sum_{i,j} \int_{C_{ij}} t\mu)\\
&= \textstyle\int_X s\mu + \int_X t\mu
\end{aligned}$$

■

We are now ready to prove one of the most important theorems in integration theory and the most useful for our purposes. This theorem is Lebesgue's monotone convergence theorem and represents perhaps the most striking difference between the Riemann integral and the Lebesgue integral. Roughly speaking it says that the integral of the limit of a sequence of functions is the limit of the integrals of the individual functions. There are several theorems like this and some, for example the dominated convergence theorem, are more general and more useful in classical analysis but the starting point of all such theorems is the monotone convergence theorem.

Theorem 3.9 (Monotone Convergence Theorem) *Let $\{f_n\}$ be a sequence of measurable functions on X and suppose that*

1. $\forall x \in X. 0 \leq f_1(x) \leq f_2(x) \leq \ldots \leq \infty$
2. $\forall x \in X. \sqcup_n f_n(x) = f(x)$.

Then f is measurable, and $\sqcup_n \int_X f_n \mu = \int_X f \mu$.

Of course we already know that f is measurable, it is only asserted here to ensure that the subsequent statement is sensible. Usually one states this with limits rather than sups, the choice is a matter of taste. We work with the complete lattice $[0, \infty]$ so that we might be dealing with functions that are not integrable but do possess a definite integral. For everywhere nonnegative functions having a definite integral is the same as being measurable.

Proof. We know that $\forall n. \int_X f_n \mu \leq \int_X f \mu$ so $K \stackrel{\text{def}}{=} \sqcup_n \int_X f_n \mu \leq \int_X f \mu$. We need to prove that $K \geq \int_X f \mu$. The latter is also a sup, $\int_X f \mu = \sqcup \int_X s \mu$ where s stands for a simple function below f. Thus if we can prove that for any simple function $s \leq f$, $\int_X s \mu \leq K$ we will be done.

What we shall prove is that for any number c in the interval $(0, 1)$ we have $c \int_X s \mu \leq K$ from which the required inequality immediately follows. Accordingly we fix such a c and a simple function s. We define the sets

$$\forall n. A_n = \{x | f_n(x) \geq c \cdot s(x)\}.$$

These are clearly measurable sets and $\forall n. A_n \subseteq A_{n+1}$ since the f_n form an increasing sequence. Now we claim that $\cup_n A_n = X$. Let $x \in X$ be arbitrary. Since $c < 1$ we have $c \cdot s(x) < f(x)$ if $f(x) > 0$. Since $\sqcup_n f_n(x) = f(x)$ there is some m such that $f_m(x) > c \cdot s(x)$ and hence $x \in A_m$. If $f(x) = 0$ then clearly $c \cdot s(x) = 0$ and $x \in A_1$. Now we clearly have

$$\int_X f_n \mu \geq \int_{A_n} f_n \mu \geq c \int_X s \mu.$$

If we take sups of both sides on the left we get K and on the right we have that $\cup_n A_n = X$ thus we have

$$K \geq c \cdot \int_X s \mu.$$

■

It is worth noting how the constant c is used to make the A_n big enough to cover X. We have the immediate corollary

Corollary 3.10 *If we have f and f_n as above and A is any measurable subset then $\int_A f\mu = \sqcup \int_A s\mu$.*

The next theorem is a basic useful fact but the real point of discussing it now is to exhibit a paradigmatic proof strategy for using the monotone convergence theorem. Let $T : (X, \Sigma_X) \longrightarrow (Y, \Sigma_Y)$ be a measurable function. A measure μ on X induces a measure ν on Y by the formula $\nu = \mu \circ T^{-1}$. If $X = Y$, $\Sigma_X = \Sigma_Y$ and $\mu = \nu$ we call T an *ergodic* transformation.

Proposition 3.11 *If we have X, Y, T as above and $f : Y \longrightarrow \mathbf{R}$ is measurable and $B \in \Sigma_Y$ then*

$$\int_{T^{-1}(B)} f \circ T\mu = \int_B f\nu$$

in the sense that if one integral exists then so does the other and they are equal.

Proof. Suppose that f is the characteristic function of some measurable set C. Then $f \circ T$ is the characteristic function of the set $T^{-1}(C)$. The left hand side of the above equation is then $\mu(T^{-1}(C) \cap T^{-1}(B))$. But this is equal to $\mu(T^{-1}(B \cap C)) = \nu(B \cap C)$ which is precisely the value of the right hand side. Since the integral is linear on simple functions and any simple function is a linear combination of characteristic functions the required equation holds for any simple function. Now consider any nonnegative function f. We have $f = \sqcup\{s | 0 \leq s \leq f\}$. The monotone convergence theorem says that $\int_B f\nu = \sqcup \int_B s\nu$, while the fact that s is simple allows us to conclude that $\int_B s\nu = \int_{T^{-1}(B)} s \circ T\mu$. Now we also note that $\sqcup s \circ T = f \circ T$. Putting all these together we get the required result. For f not nonnegative we can easily apply these arguments separately to the positive and negative parts. ∎

This proof is paradigmatic. In fact this pattern is so closely followed that one can omit the details and just invoke the pattern of this proof as a kind of "mantra". When I say "monotone convergence mantra" henceforth I mean exactly this pattern. A few exercises will familiarize any readers with this proof and then they will be able to recognize occurrences of this pattern easily. A typical application is to prove that integration is indeed linear for integrable functions, not just for simple functions.

Proposition 3.12 *If f and g are measurable functions then*

$$\int (f+g)\mu = \int f\mu + \int g\mu.$$

The proof is left as an exercise.

The most important theorem relating limits and integration is Lebesgue's dominated convergence theorem. We state this without proof.

Theorem 3.13 *Let f_n and g be integrable functions such that*

$$\forall n. |f_n| \leq g \text{ and } \lim_{n \longrightarrow \infty} f_n = f.$$

Then f is integrable and

$$\lim_{n \longrightarrow \infty} \int f_n\mu = \int f\mu.$$

This allows one to deal with sequences of functions that are not monotone increasing. However this theorem follows essentially from the monotone convergence theorem.

3.3 Riemann Integrals

In this section we discuss the relationship between the familiar Riemann integral and the Lebesgue integral. Consider real-valued functions defined on a closed interval $[a, b]$.

Definition 3.14 *We say that such a function f is Riemann-integrable with Riemann-integral α if*

$$\forall \epsilon > 0 \exists a = a_0 < a_1 < a_2 < \ldots < a_n = b. |(\sum_{i=1}^{n} f(x_i)(a_i - a_{i-1}) - \alpha| \leq \epsilon$$

whenever $x_i \in [a_{i-1}, a_i]$.

Note the way the definition is couched in terms of dividing up the domain of the function.

Example 3.15 Suppose that the interval is $[0, 1]$ and the function is $f(x) = x$. To verify that the integral is $\frac{1}{2}$ we proceed as follows. Given an ϵ we choose the $a_i = \frac{i}{n}$, so $a_i - a_{i-1} = \frac{1}{n}$. We maximize the sum by choosing $x_i = a_i$ and we minimize it by choosing $x_i = a_{i-1}$. Easy calculations then show that

$$\sum_{i=1}^{n} f(x_i)(a_i - a_{i-1}) \in [\frac{1}{2} - \frac{1}{2n}, \frac{1}{2} + \frac{1}{2n}].$$

In other words by choosing $n > \frac{2}{\epsilon}$ and dividing up the interval uniformly we get the required inequality with $\alpha = \frac{1}{2}$.

The next example is a contrast with the Lebesgue case.

Example 3.16 We again consider a function defined on $[0, 1]$. Suppose that f is 0 if x is rational and 1 if x is irrational. The Lebesgue integral of this function is 1, we shall argue that the Riemann integral does not exist. Let us choose $\epsilon = \frac{1}{2}$. Let the a_i be chosen arbitrarily. Any of the intervals $[a_{i-1}, a_i]$ will contain both rational and irrational numbers. Choosing only rationals we see that the sum is zero while choosing only irrationals the sum is 1. Thus whatever α might be there is no way of forcing the sum to be within ϵ of α for all choices of the x_i.

This example shows the role played by continuity. If f is too "wild" we cannot constrain the values of the sum by confining x to small intervals.

Theorem 3.17 *Let f be a bounded function on $[0, 1]$. The function f is Riemann-integrable iff the set of points at which f is not continuous has Lebesgue measure 0.*

Proof. We first define a function F which "measures how discontinuous f is." For a fixed $x \in [0, 1]$ we define

$$F(x) = \lim_{\epsilon \longrightarrow 0} sup_{x', x''} \{|f(x') - f(x'')| : |x - x'|, |x - x''| \leq \epsilon\}.$$

The sup is defined since f is bounded and the sup decreases as $\epsilon \longrightarrow 0$ so the limit is defined. Clearly if f is continuous at x, $F(x) = 0$. At a point of discontinuity of f, F is the value of the jump in the value of f, well-defined since f is bounded. We define $F_\delta = \{x : F(x) \geq \delta\}$. The set of all points where f is discontinuous is $D = \cup_{\delta > 0} F_\delta$.

Let us assume that f is Riemann-integrable. We claim that for all δ the set F_δ has Lebesgue measure 0. To see this we fix a δ. Now for any $\epsilon > 0$ we can find a_i as required by the definition of Riemann-integrable. Now suppose that I is the subset of $\{1, \dots, n\}$ such that that every interval $[a_{i-1}, a_i]$ with $i \in I$, contains a point of F_δ, then the entire sum in the definition of the Riemann integral can vary by at least $\delta \cdot \sum_{i \in I}(a_i - a_{i-1})$. But since f is Riemann-integrable we must have this sum bounded by $2 \cdot \epsilon$. Thus $\sum_{i \in I}(a_i - a_{i-1}) \leq \frac{2\epsilon}{\delta}$. Now F_δ, except for finitely many points (some of the a_i may be in F_δ), is contained in intervals of total length $\frac{2\epsilon}{\delta}$. Thus the outer measure of F_δ is less than $\frac{2\epsilon}{\delta}$, which is only possible for all ϵ if the outer measure, and hence the Lebesgue measure, of F_δ is 0.

Now D is contained in the union of countably many sets of measure zero

$$D \subseteq \bigcup_j F_{\frac{1}{2^j}},$$

and thus has measure 0.

To argue the other direction we use compactness. Essentially we have to show that we can construct a partition such that we can make the variation of the sum appearing in the definition of the Riemann integral as small as we please.

Recall that closed subsets of $[0, 1]$ are compact. We claim that F_δ is always closed. Suppose that we have a convergent sequence $\{x_i\}$ in F_δ and that the sequence converges to x. We need to argue that x is in F_δ. Any open neighbourhood of x contains some point x_i i.e. some point of F_δ so, in any neighbourhood of x, the variation of f is at least δ. Thus $x \in F_\delta$.

Since, by assumption, F_δ has measure 0 it can be covered by a family of open intervals of total length less than any specified ϵ. Since F_δ is compact, a finite number of these intervals suffice to cover F_δ. Without loss of generality we can take these intervals to be disjoint since the union of two overlapping open intervals is an open interval. We call these intervals $A_1 \dots, A_n$. We define $B = (\cup_i A_i)^c$, where the complement is defined with respect to $[0, 1]$; clearly B is closed hence compact. Also F takes on values less than δ on B. We can therefore find finitely many, open, disjoint intervals that cover B; we call these B intervals.

We construct a partition which bounds the variation in the sum appearing in the Riemann integral as follows. We order the end points of the A and the B intervals in order to define the partition. The intervals may overlap and we may overestimate the variation but we will bound this estimate in the right direction. The variation in the value of the sum just taken over the A-intervals is $2M\epsilon$ (where M is the bound on the value of f) since their total length is bounded by ϵ. The contribution to the variation of the sum coming from the B-intervals is δ. The total length of the B intervals may be 1 but the variation cannot be more than δ. Thus the overall bound we get is $\delta + 2M\epsilon$. Since we can choose both δ and ϵ to be arbitrarily small we can make the variation in the sum appearing in the definition of the Riemann integral arbitrarily small. ∎

The Riemann integral and the Lebesgue integral are intimately related. As the above theorem shows Riemann-integrable functions are characterized by Lebesgue measure properties of the discontinuities. A more intimate tie-up is revealed by the Riesz representation theorem. This very important theorem is beyond the scope of these notes but we can indicate the ideas. The space of continuous functions on a closed interval forms a topological vector space[1]. The vector space

[1] It also forms an algebra and various other things but we concentrate on the TVS structure for now.

structure is evident, the topological structure comes from the sup norm. Now Riesz's theorem says that any *continuous, linear* functional, say Λ, on this space can be represented by a measure μ, in other words one can find a measure μ such that

$$\Lambda(f) = \int f\mu.$$

For example the functional that evaluates a function at a point is represented by the Dirac measure. All continuous functions are Riemann-integrable so the Riemann integral defines a continuous linear functional. The measure that corresponds to this functional is precisely Lebesgue measure. In many books Lebesgue measure is first constructed this way.

3.4 Multiple Integrals

A very basic result in integration is the ability to treat double integrals as iterated integrals and to change the order of integration of the latter. The basic result, which says that under very general conditions one can do all this, is called Fubini's theorem. In order to treat this we need a little more measure theory. We see that some of these basic facts are proved by using integration, in particular the monotone convergence theorem.

We first need to define measures on product spaces. Given two measure spaces (X, Σ_X, μ) and (Y, Σ_Y, ν) we want to define a measure on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$. We call a set of the form $A \times B$, with $A \in \Sigma_X, B \in \Sigma_Y$, a *measurable rectangle*. We can define a measure ρ on $\Sigma_X \otimes \Sigma_Y$ by starting with $\rho(A \times B) = \mu(A) \cdot \nu(B)$. We use our usual conventions for arithmetic with ∞. Now in order to define a measure we need to know that ρ is countably additive.

Proposition 3.18 *ρ is countably additive on the collection of rectangles.*

Proof. Suppose that we have a countable pairwise disjoint family of rectangles $A_n \times B_n$ such that there exist A and B with $A \times B \cup_n A_n \times B_n$.

We work with characteristic functions, thus we have

$$\chi_A \chi_B = \sum_n \chi_{A_n} \chi_{B_n}.$$

Now for fixed x we have on each side a measurable function of y. Integrating wrt y using the measure ν gives the equation

$$\chi_A(x)\nu(B) = \sum_n \chi_{A_n}(x)\nu(B_n).$$

This is immediate from linearity of integration. Now we can integrate with respect to x using μ to get

$$\mu(A)\nu(B) = \sum_n \mu(A_n)\nu(B_n)$$

where we have used linearity of the integral and the monotone convergence theorem in the usual way. ∎

The rectangles form a semi-ring thus the extension theorem 2.32 says that ρ extends to a measure on the σ-field generated by the rectangles. However, this extension is not unique. The key property needed to guarantee uniqueness is σ-finiteness; i.e. the measure space is the countable union of sets of finite measure. In probability theory this condition is usually satisfied.

We are now ready to show that we have a *unique* product measure when the measures in question are σ-finite. It turns out useful to work with monotone classes. We follow the general pattern of the monotone convergence mantra. We begin by showing that we can invert the order of integration for characteristic functions.

Proposition 3.19 *Suppose that (X, Σ_X, μ) and (Y, Σ_Y, ν) be measure spaces such that $\mu(X) < \infty$ and $\nu(Y) < \infty$. Let*

$$\mathcal{F} \stackrel{\text{def}}{=} \{E \subseteq X \times Y : \int [\int \chi_E(x,y)\mu]\nu = \int [\int \chi_E(x,y)\nu]\mu\}.$$

Then $\Sigma_X \otimes \Sigma_Y \subseteq \mathcal{F}$.

Proof. We start with just rectangles and then use the fact that the smallest monotone class containing all the rectangles is just the same as the σ-field generated by the rectangles. Let $E = A \times B$ be a rectangle. Now we have $\chi_E = \chi_A \cdot \chi_B$ so

$$\int [\int \chi_E \mu]\nu = \int [\int \chi_A \chi_B \mu]\nu = \mu(A)\nu(B) = \int [\int \chi_E \nu]\mu.$$

Thus any rectangle is in $\mathcal{F}$. Clearly any finite disjoint union of rectangles is also in $\mathcal{F}$.

Now in order to show that $\mathcal{F}$ is closed under increasing and decreasing monotone limits we proceed by induction on the stages in the construction of $\mathcal{F}$ starting from the rectangles. We have just established the base case. For the inductive case we suppose that $E_n \uparrow E$ in $\mathcal{F}$ and that we can reverse the order of integration for the characteristic functions of the E_n. Now if $E_n \uparrow E$ we have $\sqcup_n \chi_{E_n} = \chi_E$. Thus by using the monotone convergence theorem we have $\sqcup_n \int \int \chi_{E_n} \mu\nu = \int \int \chi_E \mu\nu$ and also $\sqcup_n \int \int \chi_{E_n} \nu\mu = \int \int \chi_E \nu\mu$. By assumption $\int \int \chi_{E_n} \mu\nu = \int \int \chi_{E_n} \nu\mu$ for all n, so we get $\int \int \chi_E \mu\nu = \int \int \chi_E \nu\mu$. Thus $E \in \mathcal{F}$. If we have $E_n \downarrow E$ we proceed in exactly the same way except that we use the dominated convergence theorem rather than the monotone convergence theorem. It is here that we need the finiteness assumption on the measure. Thus $\mathcal{F}$ is a monotone class. Note that we are always dealing with measurable functions at every stage of this proof. ∎

Theorem 3.20 *Let (X, Σ_X, μ) and (Y, Σ_Y, ν) be σ-finite measure spaces. Then ρ extends uniquely to a measure on $\Sigma_X \otimes \Sigma_Y$ such that for all $E \in \Sigma_x \otimes \Sigma_Y$*

$$\rho(E) = \int [\int \chi_E(x,y)\mu]\nu = \int [\int \chi_E(x,y)\nu]\mu.$$

Proof. We start by assuming that the measures μ and ν are finite. Let $\eta(E) = \int [\int \chi_E(x,y)\mu]\nu$ for $E \in \Sigma_X \otimes \Sigma_Y$. Now this integral is defined since the measures are finite, and, as we have seen, the order of integration can be reversed. The set function η is finitely additive and by using monotone convergence theorem it is also countably additive; thus it is a measure. Suppose that we have another measure ξ with the same properties. By the same arguments as before we can prove that the sets for which ξ and η agree include all finite unions of disjoint rectangles and form a monotone class and hence includes all of $\Sigma_X \otimes \Sigma_Y$. Thus we have uniqueness.

Now we extend the result to the σ-finite case. Let X_n and Y_m satisfy, $\cup_n X_n = X$, $\forall n.\mu(X_n) < \infty$, $\cup_m Y_m = Y$ and $\forall m.\nu(Y_m) < \infty$. Let $E \in \Sigma_X \otimes \Sigma_Y$ and define sets $E_{n,m}$ by $E_{n,m} = E \cap (X_n \times Y_m)$. Now for these sets we have

$$\int\int \chi_{E_{n,m}} \mu\nu = \int\int \chi_{E_{n,m}} \nu\mu.$$

We can sum all these terms and use monotone convergence to conclude that

$$\int\int \chi_E \mu\nu = \int\int \chi_E \nu\mu.$$

Thus we can define η as before. To establish that it is a measure we use countable additivity of μ and ν to show that η is countably additive. Uniqueness follows from the fact that any other measure must agree on the $E_{n,m}$ according to the previous paragraph and hence on E by monotone convergence. ∎

The measure whose existence is established above is called the product measure and is written $\mu \times \nu$. What we have done so far is the first step in the chain of reasoning that one needs to invoke the monotone convergence theorem. In other words we have shown that we can change the order of integration if we are integrating characteristic functions. We can now complete the argument in the familiar way and prove

Theorem 3.21 (Fubini-Tonelli) *Let (X, Σ_X, μ) and (Y, Σ_Y, ν) be σ-finite measure spaces and let f be a measurable function with respect to $\Sigma_X \otimes \Sigma_Y$ from $X \times Y$ to $[0, \infty]$. Then*

$$\int f(\mu \times \nu) = \int [\int f\mu]\nu = \int [\int f\nu]\mu.$$

The same holds for not necessarily positive integrable functions.

The integral $\int f\mu$ is defined for ν-almost all y[2] and the integral $\int f\nu$ is defined for μ-almost all x. This will not affect the values of the integrals.

Proof. For characteristic functions this is just the last theorem. For simple functions we have the result by linearity of integration. For arbitrary positive f it follows from the monotone convergence theorem and the usual case analysis for the positive and negative parts gives the result for the case of integrable functions. ∎

We have stated these theorems for binary products but by an easy induction on n it holds for products of n σ-finite spaces as well. Constructing measures on larger products is more subtle but is important for the analysis of stochastic processes.

[2]In other words, the set on which the integral is not defined has ν-measure 0.

3.5 Exercises

Unless otherwise stated these exercises refer to a fixed measure space (X, Σ, μ) and the phrase "measurable function", without further qualification means "measurable real-valued function."

1. Show that the integral is a linear function of measurable functions, i.e. for f, g measurable show that
$$\int (f+g)\mu = \int f\mu + \int g\mu.$$

2. Show that if f_n is a family of nonnegative, real-valued, measurable functions then $f \stackrel{\text{def}}{=} \sum_n f_n$ is measurable and that
$$\int f\mu = \lim_{n \longrightarrow \infty} \int f_n \mu.$$

3. Let A_1 be $[0,1]$, $A_2 = A_1 \cap (\frac{1}{3}, \frac{2}{3})^c$, $A_3 = A_2 \cap (\frac{1}{9}, \frac{2}{9})^c \cap (\frac{7}{9}, \frac{8}{9})^c$ and so on. We have $A \stackrel{\text{def}}{=} \cap_i A_i$ is the Cantor set. Now we define a function $f : [0,1] \longrightarrow [0,1]$ by
$$f(x) = \begin{cases} 1 & \text{if } x \in A \\ \frac{1}{2} & \text{if } x \in A_1 \cap A_2^c \\ \frac{1}{4} & \text{if } x \in A_2 \cap A_3^c \end{cases}$$
Show that this function is Riemann integrable and calculate its Riemann integral. [Hint the answer is $\frac{1}{4}$.]

4. The monotone convergence theorem is not true for arbitrary directed sets of measurable functions. Construct an example showing this. [Hint: Consider the interval $[0,1]$ and the characteristic functions of all finite subsets of this interval.] (I am indebted to Jørgen Hoffman-Jørensen for pointing this out to me.)

Bibliography

[Ash72] R. B. Ash. *Real Analysis and Probability.* Academic Press, 1972.

[Bil95] P. Billingsley. *Probability and Measure.* Wiley-Interscience, 1995.

[Bre68] L. Breiman. *Probability.* Addison-Wesley, 1968.

[Ger85] Robert Geroch. *Mathematical Physics.* Chicago Lectures in Physics. University of Chicago Press, 1985.

[Jec71] Thomas J. Jech. *Lectures in Set Theory.* Number 217 in Lecture Notes In Mathematics. Springer-Verlag, 1971.

[KT66] J. F. C. Kingman and S. J. Taylor. *Introduction to Measure and Probability.* Cambridge University Press, 1966.

[Mal95] P. Malliavin. *Integration and Probability.* Number 157 in Graduate Texts in Mathematics. Springer-Verlag, 1995. French edition appeared in 1993.

[Rud66] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1966.

[Sol70] R. M. Solovay. A model of set theory in which every set of reals is lebesgue measurable. *Annals of Mathematics*, 92:1–56, 1970.