

# Chapter 1

## Probability Spaces and Independence

This chapter has two goals. The first is to establish the mathematical foundations of probability. Probability theory is carried out within the framework of probability spaces, which are just special examples of measure spaces. Section 1.1 explains how probability spaces are defined and presents the basic measure-theoretic tools needed to construct and analyze them. The second goal is to introduce the concept of independence. Independence is absolutely central to all aspects of probability and stochastic processes, and is largely responsible for the distinct flavor of probability theory as a branch of analysis. The two themes—measure theory and independence—are intertwined in the examples discussed in this chapter. In fact, a centerpiece of the chapter is a careful construction of a probability space modeling an infinite sequence of independent coin tosses. This example is one of the simplest probability spaces requiring a non-trivial construction; at the same time, it typifies the kind of “infinite-dimensional” measure space characteristic of probability theory.

Probability theory is a subject motivated by applications, and one of its attractions is the rich class of problems that can be understood in detail by an elegant mix of combinatorics and analysis. Another is the derivation of interesting limit laws, such as the law of large numbers or the central limit theorem, that have a universal character, because they are independent of particular distributional assumptions. The student drawn to probability for these reasons will have to be patient in reading this first chapter. It is about foundations, is mostly abstract, and does not enter into even mildly interesting calculations. Computing probabilities of the outcome of a finite number of independent coin

tosses will be about as complicated as it gets! And even for this we will present only general formulas. However, the measure theory background is essential; it supplies both the scaffolding for a mathematically rigorous theory and also powerful analytic tools that are used repeatedly. It is especially important to stochastic process theory, a study of which is begun in these notes. So it pays to master this foundational material well at the start.

## 1.1 Probability spaces

### 1.1.1 A few comments on the meaning of probability.

Here is a statement you would certainly believe: the probability of getting an even number in one roll of a die is  $1/2$ . Here is another you could verify by a simple combinatorial argument: the probability of getting exactly two heads in 3 tosses of a fair coin is  $3/8$ . We understand these statements intuitively, but what do they mean precisely? More generally, what does it mean to assign a probability to an uncertain event? As a guide to intuition, it is useful to make a few remarks about interpreting probabilities, before delving into the mathematical theory. Be warned, however, that this is a philosophical issue, meaning it admits contending approaches and has no, single, scientifically verifiable resolution. Be warned also that that this discussion is philosophically naive and simplistic, reflecting, I am afraid, the philosophical naiveté of the author!

Most probabilists and statisticians subscribe to one of two approaches to the meaning of probability. One is called the *frequentist interpretation*, and it applies most convincingly to random phenomena that can, in principle, be observed as often as desired under identical circumstances. A coin toss, a die roll, or, more generally, any game of chance provide typical examples. Another example is Brownian motion, the small scale, random motion, due to molecular collisions, of a microscopic particle suspended in a fluid or gas. Imagine that we observe  $N$  instances of such a phenomenon, and that the outcomes of the different instances do not influence each other. If  $A$  is some event concerning the outcome, the number of times  $A$  occurs divided by  $N$  is called the empirical frequency of  $A$  in those  $N$  trials; denote it by  $f_N^A$ . Of course, for any given  $N$ ,  $f_N^A$  is itself random: different runs of  $N$  trials will typically yield different values of  $f_N^A$ . However, the frequentist believes that  $f_N^A$  will tend to a limiting value as  $N \rightarrow \infty$ , and that *this limit will not be random, but will be the same for any sequence of trials*. For the frequentist, the probability of  $A$  is precisely this limiting frequency. For example, saying that the probability of getting an even

number in a die roll is  $1/2$  means that in a long run of die rolls, approximately  $1/2$  are even, and that this approximation becomes exact as the number of rolls increases toward infinity, in *any* infinite sequence of independent trials. The frequentist definition of probability does make good intuitive sense; if the long run empirical frequency of event  $A$  is greater than that of event  $B$ ,  $A$  occurs more frequently on average and hence is “more probable.” Since the empirically observed frequencies of different possible outcomes reflect only physical laws governing the experimental set-up, the frequentist’s probability is an objective quantity independent of the observer.

The frequentist approach does not apply, however, to situations such as the next presidential election, or tomorrow’s horse race, that are unique and not repeatable. Yet the outcomes are still uncertain and people still assign them probabilities, if only to make book for betting. In such cases, people form judgments about the relative likelihood of different outcomes using knowledge of previous performance (whether of politicians or horses), polls, recent history, the alignment of the planets, or whatever else they deem relevant. Necessarily, these judgments will be subjective and vary from individual to individual; an expert in presidential politics will have a much more precise and informed assessment about the presidential election than a casual observer, but even experts will disagree among themselves. The probabilities people assign to outcomes are thus subjective assessments of relative likelihood, subject to change as they acquire new data. The so-called *Bayesian* philosophy insists that the only coherent interpretation of probabilities is to regard them in all cases as subjective measures of likelihood, even in those circumstances, such as games of chance, where the frequentist approach seems natural. (After all, it is not really possible to conduct an infinite number of independent trials.) It is a very statistical outlook, since it brings to front and center the problem of updating one’s subjective probabilities as new observations and information become available. Indeed, the term *Bayesian* comes from Bayes’ rule for conditional probabilities, which is the mathematical tool for updating probabilities. (I believe that Mr. Bayes (1702-61), the inventor of this rule, was not himself a Bayesian; indeed, he predates the formulation of the Bayesian philosophy.)

There are other viewpoints, as well. In fact, the early developers of probability were neither Bayesian nor frequentists. They thought that all assignments of probabilities should be based on reducing random phenomena to a set of cases that must, by their nature, be considered equally probable. This idea is found, for instance, in Laplace’s work in probability. It makes sense for analyzing simple games of chance, as the early pioneers of probability tried to do. For example,

in computing the probabilities of different card hands, it is reasonable, assuming adequate shuffling, to treat each possible hand as equally probable. As a general principle, seeking equally probable basic outcomes has the virtue of determining probabilities objectively, but without reference to a limiting frequency. However, especially for more complicated phenomenon, it begs the question of when a decomposition into equally probable cases has been achieved, and it is not clear how to apply the idea in models with an infinite number of possible outcomes.

The interpretation of probability is a subject of lively philosophical debate and it does have practical consequence for statistical practice. Fortunately, it is not necessary to take sides in the debate, nor even to pursue the different interpretations in any depth, in order to formulate a mathematical theory of probability. Both Bayesians and frequentists agree on a minimal set of rules constraining the assignment of probabilities. These rules, which are axiomatized in the notion of a probability space, are the basis of the theory.

### 1.1.2 Probability Spaces

We will approach the definition of a probability space from the viewpoint of modeling. Random behavior occurs in widely diverse settings and for a variety of reasons. For a comprehensive theory not tied to any particular application, one would like a common mathematical framework for building probabilistic models. This is the purpose of the probability space concept. It serves as a universal template for probabilistic modeling.

The abstract notion of a probability space provides two things: a common mathematical language for probability modeling, and axioms that constrain how probabilities are assigned. The language is easy to describe and very natural. Consider a random trial and let  $\Omega$  be the set listing all its possible outcomes. The purpose of a model is to assign probabilities to different “events” concerning the actual outcome. The idea of an “event” is informal, but in practice it always ultimately takes the form: “the outcome falls in a certain subset of  $\Omega$ .” For example, the event a single die roll comes up even is just the probability that the outcome of the roll is in the subset  $\{2, 4, 6\}$  of  $\{1, 2, 3, 4, 5, 6\}$ . Therefore, it makes sense to *identify* events with subsets of  $\Omega$ . If  $A \subset \Omega$ , we say *event A occurs* if the outcome falls in  $A$ . Formulating a probability model then means assigning numbers  $\mathbb{P}(A)$  to subsets of  $\Omega$ ,  $\mathbb{P}(A)$  being the probability that  $A$  occurs. This will be the general approach: a probability space will feature a function  $\mathbb{P}$ , called a probability measure, defined on subsets, called events, of a set  $\Omega$ .

The “axiom” part of the definition of probability space stipulates a few,

basic properties, consistent with intuitive notions of probability, that  $\mathbb{P}$  must satisfy. The simplest obvious requirements are that  $\mathbb{P}(A)$  be non-negative for any event  $A$  and that  $\mathbb{P}(\Omega) = 1$ , the latter merely expressing the conventions that  $\Omega$  includes all possible outcomes and that an event with probability 1 is certain to occur. The final and more important property is additivity; if  $A$  and  $B$  are disjoint events, then  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ . Additivity is natural to the idea of probability as a measure of likelihood. In fact, the frequentist interpretation demands additivity, because as the number of trials tend to infinity, the limiting empirical frequency of  $A \cup B$  is the sum of the individual limiting frequencies, as long as  $A$  and  $B$  are disjoint. Later we will strengthen additivity to countable additivity.

It is also necessary to place conditions on the domain of  $\mathbb{P}$ , the class of those subsets of  $\Omega$  that qualify as events. This is not immediately obvious. Why not always let the domain be the power set  $2^\Omega$ , the set of all subsets of  $\Omega$ ? Certainly, this is the natural thing to do whenever  $\Omega$  is finite or countable. However, if  $\Omega$  is uncountable, and if we use standard set theory with the axiom of choice, the power set is too rich. The mathematical techniques for establishing existence of countably additive measures do not, in general, yield the power set as a domain of  $\mathbb{P}$ . And in some cases, it is not even possible to define  $\mathbb{P}$  on all subsets and maintain countable additivity. However, if the class of events must be restricted, then we at least want it to be closed with respect to elementary operations of union, intersection, and taking complements, since these are used all the time in even elementary probability calculations. The following important definitions formalize these closure properties.

**Definition 1** *Let  $\Omega$  be a non-empty set. Let  $\mathcal{F}$  be a non-empty collection of subsets of  $\Omega$ .*

(A).  $\mathcal{F}$  is an algebra if it is closed under complements and finite unions and intersections. That is,

- (i)  $A \in \mathcal{F}$  implies  $A^c \in \mathcal{F}$ ;
- (ii) for any finite  $n$ ,  $A_1, \dots, A_n \in \mathcal{F}$  implies  $A_1 \cup \dots \cup A_n \in \mathcal{F}$ ;
- (iii) for any finite  $n$ ,  $A_1, \dots, A_n \in \mathcal{F}$  implies  $A_1 \cap \dots \cap A_n \in \mathcal{F}$ .

(B).  $\mathcal{F}$  is a  $\sigma$ -algebra if it is closed under complements and countable unions and intersections. That is,  $\mathcal{F}$  satisfies (i) and

- (ii')  $A_1, A_2, \dots \in \mathcal{F}$  implies  $\cup_1^\infty A_i \in \mathcal{F}$ ;

(iii')  $A_1, A_2, \dots \in \mathcal{F}$  implies  $\bigcap_1^\infty A_i \in \mathcal{F}$ .

Note that if  $\mathcal{F}$  is an algebra or  $\sigma$ -algebra of subsets of  $\Omega$ , then

$$\Omega \in \mathcal{F} \quad \text{and} \quad \emptyset \in \mathcal{F}.$$

Indeed, since  $\mathcal{F}$  is non-empty, there is at least one  $A$  in  $\mathcal{F}$ . But then  $A^c \in \mathcal{F}$  and then also  $\Omega = A \cup A^c$  and  $\emptyset = \Omega^c$  are in  $\mathcal{F}$ .

The conditions defining algebras and  $\sigma$ -algebras in Definition 1 are partly redundant. The reader should prove the following simple result, which requires only simple applications of De Morgan's laws.

**Lemma 1** *Let  $\mathcal{F}$  be a non-empty collection of subsets of  $\Omega$ . If  $\mathcal{F}$  satisfies just conditions (i) and (ii) or just (i) and (iii) it is an algebra. If  $\mathcal{F}$  satisfies just conditions (i) and (ii') or just (i) and (iii') it is a  $\sigma$ -algebra.*

A simple argument by mathematical induction also shows that to establish (ii) or (iii) in checking whether  $\mathcal{F}$  is an algebra, it suffices to check the case  $n = 2$ .

We are finally in a position to define probability spaces.

**Definition 2** *The triple  $(\Omega, \mathcal{F}, \mathbb{P})$  is called a finitely additive probability space if  $\Omega$  is a non-empty set,  $\mathcal{F}$  is an algebra of subsets of  $\Omega$ , and  $\mathbb{P}$  is a non-negative function on  $\mathcal{F}$  satisfying  $\mathbb{P}(\Omega) = 1$  and*

$$\mathbb{P}\left(\bigcup_1^n A_i\right) = \sum_1^n \mathbb{P}(A_i) \quad \text{for disjoint } A_1, \dots, A_n \text{ in } \mathcal{F}. \quad (1.1)$$

$\mathbb{P}$  is then called a finitely additive probability measure and the sets in  $\mathcal{F}$  are called events.

**Example 1.1.** *Probability space for the roll of a fair die.* Let  $\Omega := \{1, 2, 3, 4, 5, 6\}$ , let  $\mathcal{F}$  be the collection of all subsets of  $\Omega$ , and let  $\mathbb{P}(A) := |A|/6$ , for every  $A \subset \Omega$ , where  $|A|$  denotes the cardinality of  $A$ . This is a probability model for one roll of a fair die, because for each  $i \in \{1, \dots, 6\}$ ,  $\mathbb{P}(\{i\}) = 1/6$ . It is easy to show that  $\mathbb{P}$  is finitely additive.  $\diamond$

**Example 1.2.** *Uniform distribution on  $[0, 1]$ , finitely additive version.* Imagine an experiment whose outcome is a number in  $[0, 1]$  and is equally likely to fall anywhere in the interval. The outcome space is  $[0, 1]$ . Let  $\mathcal{F}$  be the collection of all finite disjoint unions of subintervals of  $[0, 1]$ . At a minimum, we would

certainly like any subinterval of  $[0, 1]$ , whether closed, open, or half closed-half open, to be an events. Thus the class of events should include  $\mathcal{F}$ . But it turns out that  $\mathcal{F}$  is in fact an algebra. The proof is left as an exercise, which is easy using a technical lemma to be introduced in the following section.

Since there is no region of the  $[0, 1]$  more likely than any other, the probability of an interval should depend only on its length, not its position. By additivity, the probability must be linear as a function of length, and since  $\mathbb{P}([0, 1]) = 1$ , we define

$$\mathbb{P}([a, b]) = \mathbb{P}((a, b]) = \mathbb{P}([a, b)) = \mathbb{P}((a, b)) = b - a,$$

for all  $0 \leq a \leq b \leq 1$ . The probability of any event in  $\mathcal{F}$  that are finite disjoint unions of more than one interval are determined from this formula by finite additivity. Clearly, for any  $A \in \mathcal{F}$ ,  $\mathbb{P}(A)$  is just the total length of  $A$ .  $\diamond$

Example 1.2 has an annoying feature it shares with most probability spaces on uncountable outcome spaces. Although the elements  $\omega$  of  $\Omega$  are the possible outcomes,  $\mathbb{P}(\{\omega\}) = 0$  for every  $\omega$ . So every time the experiment runs, an outcome of probability zero occurs! This paradox does not mean the uniform distribution is useless. In real life, no measurement is arbitrarily accurate. If measurement are accurate only to  $1/N$ , the experiment is really returning one of the numbers in the finite set  $\{0, 1/N, 2/N, \dots, 1\}$  and a uniform distribution means each outcome has the same probability, namely,  $1/(N + 1)$ . When  $N$  is large, the uniform distribution on  $[0, 1]$  will be a good approximation to the discrete model, and one that is easier to work with.

Much of probability theory is concerned with questions about limits and to discuss limits requires operations on countable sets of events. Finite additivity is not a powerful enough axiom for analyzing limits, but countable additivity is.

**Definition 3** *The triple  $(\Omega, \mathcal{F}, \mathbb{P})$  is called a probability space if  $\Omega$  is a non-empty set,  $\mathcal{F}$  is a  $\sigma$ -algebra of subsets of  $\Omega$ , and  $\mathbb{P}$  is a non-negative function on  $\mathcal{F}$  satisfying  $\mathbb{P}(\Omega) = 1$  and*

$$\mathbb{P}\left(\bigcup_1^{\infty} A_i\right) = \sum_1^{\infty} \mathbb{P}(A_i) \quad \text{for any countable disjoint family } A_1, A_2, \dots \text{ of subsets in } \mathcal{F}. \quad (1.2)$$

Definition 3 is the standard probability space used in probability theory, which is why we refer to them simply as “probability spaces,” without explicit

mention of countable additivity. All the theory discussed in these notes assumes countable additivity. We have defined finitely additive probability spaces for two reasons. First, a probability space are usually constructed in practice by defining a finitely additive probability measure on an algebra, and then extending it using theorems from measure theory. Second, there are applications which are modeled by finitely additive probability spaces that cannot be extended to countably additive spaces. Note that any probability space is a finitely additive probability space.

**Example 1.3.** *Discrete probability spaces.* Let the outcome space  $\Omega$  be a countable or finite set. Let  $\mathcal{F}$  be the power set of  $\Omega$ . Suppose that for each  $\omega \in \Omega$ , we have a number  $p_\omega \geq 0$  and that

$$\sum_{\omega \in \Omega} p_\omega = 1.$$

Define  $\mathbb{P}(A) = \sum_{\omega \in A} p_\omega$ , for any  $A \subset \Omega$ . It is an easy exercise to show that

$(\Omega, \mathcal{F}, \mathbb{P})$  is a probability space. In this model,  $p_\omega = \mathbb{P}(\{\omega\})$  is the probability that the outcome is  $\omega$  for each  $\omega \in \Omega$ .  $\diamond$

**Example 1.4.** *Uniform discrete probability space.* This is a particularly important example of a discrete probability space in which  $\Omega$  is finite and each outcome in  $\Omega$  is equally probable. Thus,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}, \quad A \subset \Omega.$$

where  $|A|$  denotes the cardinality of  $A$ . This is a model for a single random selection of an element of  $\Omega$  in which each element has an equal probability of being selected. Example 1.1 is a special case.

### 1.1.3 Elementary properties of probability measures

In this section we develop some immediate and elementary consequences of the axioms of probability.

**Theorem 1** *Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a finitely additive probability space. Then (i)  $\mathbb{P}(\emptyset) = 0$ ; (ii) if  $A \in \mathcal{F}$ ,  $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ ; (iii) if  $A \subset B$  and  $A, B \in \mathcal{F}$ , then  $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$ ; and (iv) if  $A, B \in \mathcal{F}$ ,  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ .*

The last identity (iv) is called the inclusion-exclusion principle and it has a generalization to arbitrary finite unions: if  $A_1, \dots, A_n$  are in  $\mathcal{F}$ ,

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \sum_{r=1}^n (-1)^{r+1} \sum_{1 \leq i_1 < \dots < i_r \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r}) \quad (1.3)$$

The proof of this theorem is left as an exercise; properties (i)-(iv) are elementary and the generalized inclusion-exclusion identity can be proved by induction.

The next result derives “continuity” properties of (countably additive) probability measures.

**Theorem 2** *Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space.*

(a)  *$\mathbb{P}$  is continuous from below; that is, if  $A_n$  is an increasing sequence of events,*

$$\mathbb{P}\left(\bigcup_1^\infty A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n). \quad (1.4)$$

(b)  *$\mathbb{P}$  is continuous from above; that is, if  $A_n$  is a decreasing sequence of events,*

$$\mathbb{P}\left(\bigcap_1^\infty A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n). \quad (1.5)$$

**Proof:** (a). Let  $A_1 \subset A_2 \subset \dots$  be an increasing sequence of events. Define  $B_1 = A_1$  and for  $n \geq 2$ , define  $B_n = A_n - A_{n-1}$ . Then  $B_1, B_2, \dots$  are disjoint,  $\bigcup_1^n B_i = A_n$  for every  $n \geq 1$ , and  $\bigcup_1^\infty A_i = \bigcup_1^\infty B_i$ . Therefore, using countable additivity,

$$\mathbb{P}\left(\bigcup_1^\infty A_i\right) = \sum_1^\infty \mathbb{P}(B_i) = \lim_{n \rightarrow \infty} \sum_1^n \mathbb{P}(B_i) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

This proves (a). To prove (b), use the fact that  $\left(\bigcap_1^\infty A_n\right)^c = \bigcup_1^\infty A_n^c$ . If the sequence  $A_1, A_2, \dots$  is decreasing, then  $A_1^c, A_2^c, \dots$  is increasing. Thus from (a),

$$\mathbb{P}\left(\bigcap_1^\infty A_n\right) = 1 - \mathbb{P}\left(\bigcup_1^\infty A_n^c\right) = \lim_{n \rightarrow \infty} 1 - \mathbb{P}(A_n^c) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n). \quad \diamond$$

There is a converse to Theorem 2; combined with Theorem 1.2, it says that countable additivity and continuity from above (or below) are equivalent for finitely additive probability measures. We use the notation  $A_n \downarrow A$  to indicate that  $A_1, A_2, \dots$  is a decreasing sequence of sets and  $\bigcap_1^\infty A_n = A$ .

**Theorem 3** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a finitely additive probability measure and assume in addition that  $\mathcal{F}$  is a  $\sigma$ -algebra. Suppose that for any sequence  $\{A_n\}$  of events such that  $A_n \downarrow \emptyset$ ,  $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0$ ; (in this case we say that  $\mathbb{P}$  is continuous from above at  $\emptyset$ ). Then  $\mathbb{P}$  is countably additive.

**Proof:** Let  $B_1, \dots$  be a disjoint sequence of events. Let

$$A_n = \bigcup_1^n B_i.$$

Then  $A_n \downarrow \emptyset$  and so  $\mathbb{P}(A_n) \downarrow 0$  as  $n \rightarrow \infty$ . By finite additivity of  $\mathbb{P}$ ,

$$\mathbb{P}\left(\bigcup_1^\infty B_i\right) = \mathbb{P}\left(\bigcup_1^n B_i\right) + \mathbb{P}(A_{n+1}) = \sum_1^n \mathbb{P}(B_i) + \mathbb{P}(A_{n+1}),$$

for every  $n$ . By letting  $n \rightarrow \infty$ ,

$$\mathbb{P}\left(\bigcup_1^\infty B_i\right) = \lim_{n \rightarrow \infty} \sum_1^n \mathbb{P}(B_i) = \sum_1^\infty \mathbb{P}(B_i). \quad \diamond$$

### 1.1.4 Elementary theory of algebras and $\sigma$ -algebras

We will see over and over again that in probability theory, events naturally come bundled together in  $\sigma$ -algebra packages. In this section we discuss how  $\sigma$ -algebras are generated from more elementary collections of subsets. We start first with how algebras are constructed, as this is a common stepping stone on the way to constructing  $\sigma$ -algebras, when defining probability spaces in practice.

Let  $\mathcal{E}$  be a collection of subsets of a nonempty outcome space  $\Omega$ . We want to build an algebra of subsets containing  $\mathcal{E}$ . This is particularly simple if  $\mathcal{E}$  consists of the subsets  $\{V_1, \dots, V_n\}$  of a finite partition of  $\Omega$ . Then the collection of all finite unions of subcollections of  $\{V_1, \dots, V_n\}$  is an algebra. For arbitrary  $\mathcal{E}$ , we have to work a little harder, but not much. Define the successive collections,

$$\begin{aligned} \mathcal{E}_1 &\triangleq \{A; A \in \mathcal{E} \text{ or } A^c \in \mathcal{E}\}, \\ \mathcal{E}_2 &\triangleq \{A; A \text{ is a finite intersection of sets in } \mathcal{E}_1\}, \\ \mathcal{E}_3 &\triangleq \{A; A \text{ is a finite disjoint union of sets in } \mathcal{E}_2\}. \end{aligned}$$

**Theorem 4**  $\mathcal{E}_3$  is the smallest algebra that contains  $\mathcal{E}$ .

**Proof:** Clearly any algebra that contains  $\mathcal{E}$  must contain  $\mathcal{E}_3$ . It remains to show that  $\mathcal{E}_3$  is an algebra.

Consider first the case in which  $\mathcal{E} = \{A_1, \dots, A_n\}$  is finite. For  $i = 0, 1$ , let

$$A_k^{(i)} \triangleq \begin{cases} A_k, & \text{if } i = 0; \\ A_k^c & \text{if } i = 1. \end{cases}$$

For each  $\rho = (\rho_1, \dots, \rho_n) \in \{0, 1\}^n$ , define

$$V_\rho \triangleq A_1^{(\rho_1)} \cap \dots \cap A_n^{(\rho_n)}$$

The family  $\{V_\rho; \rho \in \{0, 1\}^n\}$  is contained in  $\mathcal{E}_2$  and its members form a disjoint partition of  $\Omega$ , as is easily checked. Therefore the family  $\mathcal{A}$  of finite disjoint unions of sets in  $\{V_\rho; \rho \in \{0, 1\}^n\}$  is an algebra. Clearly  $\mathcal{A} \subset \mathcal{E}_3$ . To finish the proof, it suffices to show that  $\mathcal{E} \subset \mathcal{A}$ , because then  $\mathcal{A}$  is an algebra containing  $\mathcal{E}$ . However, this is easy, because for each  $i$ ,

$$A_i = \bigcup_{\rho, \rho_i=0} V_\rho.$$

For a general family  $\mathcal{E}$ , let

$$\mathcal{A} \triangleq \bigcup_{\mathcal{C} \subset \mathcal{E}, |\mathcal{C}| < \infty} \mathcal{C}$$

This is clearly contained in  $\mathcal{E}_3$ . It is left to the reader to show that  $\mathcal{A}$  is an algebra.  $\diamond$

From this result we can derive another useful criterion for building an algebra.

**Theorem 5** *Let  $\mathcal{E}$  be non-empty family of subsets of  $\Omega$  satisfying (i) if  $A, B \in \mathcal{E}$ , then so is  $A \cap B$ ; (ii) if  $A \in \mathcal{E}$ , then  $A^c$  is a finite disjoint union of sets in  $\mathcal{E}$ . Then the family  $\mathcal{A}$  consisting of all finite disjoint unions of elements of  $\mathcal{E}$  is an algebra.*

**Proof:** Let  $\mathcal{E}_i$ ,  $i = 1, 2, 3$  be defined as above. Certainly,  $\mathcal{A} \subset \mathcal{E}_3$  and so, to finish the proof, it remains only to show that  $\mathcal{E}_3 \subset \mathcal{A}$ . Assumption (ii) on  $\mathcal{E}$  implies  $\mathcal{E}_1 \subset \mathcal{A}$ . Now,  $\mathcal{A}$  is closed under finite intersections. Indeed, if  $A = \cup_1^n A_i$  and  $B = \cup_1^m B_j$  where  $A_1, \dots, A_n$  are disjoint and  $B_1, \dots, B_m$  are disjoint, then  $A \cap B = \cup_{i=1}^n \cup_{j=1}^m A_i \cap B_j$ , which, because of assumption (i), is again a finite disjoint union of elements of  $\mathcal{E}$ . It follows that  $\mathcal{E}_2$  is also contained in  $\mathcal{A}$ . But  $\mathcal{A}$

is also clearly closed under finite disjoint unions. Hence  $\mathcal{E}_3$  is also contained in  $\mathcal{A}$ .  $\diamond$

**Example 1.5.** Let  $\mathcal{E}$  be the collection of all subintervals of  $\mathbb{R}$  of the form  $(a, b]$ ,  $(-\infty, b]$ ,  $(a, \infty)$ , or  $(-\infty, \infty)$ . This family satisfies properties (i) and (ii) of Theorem 5. Hence the collection of all finite disjoint unions of intervals in  $\mathcal{E}$  is an algebra of subsets of  $\mathbb{R}$ .

Similarly, the family of all subintervals, whether they be closed, open, or half-closed/half-open, satisfies (i) and (ii) of Theorem 5. Thus the set of all finite disjoint unions of subintervals of  $[0, 1]$  is an algebra. (See Example 1.2.)

Unfortunately,  $\sigma$ -algebras cannot be built so simply starting from an arbitrary family  $\mathcal{E}$ . Define  $\mathcal{E}_1$  as above, then let  $\mathcal{E}_2$  be the family of all countable unions of sets in  $\mathcal{E}$  or complements thereof. In general,  $\mathcal{E}_2$  will not be a  $\sigma$ -algebra, so one has to continue taking countable unions and their complements. For  $n \geq 0$  define  $\mathcal{E}_n$  recursively so that it is the family of all countable unions of sets in  $\mathcal{E}_{n-1}$  or their complements. It can happen that no  $\mathcal{E}_n$  is a  $\sigma$ -algebra, nor is  $\cup_1^\infty \mathcal{E}_n$ . To obtain a  $\sigma$ -algebra constructively in general, one must extend this recursive procedure to define a family of collections  $\{\mathcal{E}_\alpha\}$  indexed by the countable ordinals and take the union of  $\mathcal{E}_\alpha$  over all countable ordinals. The reader may find a more detailed discussion at the end of Chapter 1 in G.B. Folland, *Real Analysis: Modern techniques and their application*, Wiley Interscience, second edition, New York, 1999. We will not need this construction. We will get around the problem using the following observation. Let  $\{\mathcal{F}_\beta; \beta \in I\}$  be any family of  $\sigma$ -algebras of subsets of a set  $\Omega$ . Then

$$\bigcap_{\beta \in I} \mathcal{F}_\beta$$

is a  $\sigma$ -algebra. The elementary derivation of this result is left to the reader. Now let  $\mathcal{E}$  be a family of subsets of  $\Omega$ . Define,

$$W \triangleq \{\mathcal{F}; \mathcal{E} \subset \mathcal{F}, \mathcal{F} \text{ is a } \sigma\text{-algebra of subsets of } \Omega\}$$

$W$  is never empty, because it always contains at least the power set of  $\Omega$ . Thus

$$\sigma(\mathcal{E}) \triangleq \bigcap_{\mathcal{F} \in W} \mathcal{F}.$$

defines the smallest  $\sigma$ -algebra containing  $\mathcal{E}$ . It is called the  $\sigma$ -algebra generated by  $\mathcal{E}$ .

In practice, most uncountable outcome spaces encountered in probability theory have topological structure. If  $\Omega$  is a topological space, we define the *Borel* sets of  $\Omega$  to be the sigma algebra generated by the open sets. The Borel sets of  $\mathbb{R}^n$  will be denoted by  $\mathcal{B}(\mathbb{R}^n)$ .

New  $\sigma$ -algebras of events are often generated by “pullbacks” of  $\sigma$ -algebras that have already been defined. We state the construction as a lemma, whose proof is left to the reader.

**Lemma 2** *Let  $\mathcal{F}$  be a  $\sigma$ -algebra of subsets of  $\Omega$ . Let  $\Theta$  be a non-empty set and let  $f : \Theta \rightarrow \Omega$ . Then  $f^{-1}(\mathcal{F}) \triangleq \{f^{-1}(U); U \in \mathcal{F}\}$  is a  $\sigma$ -algebra of subsets of  $\Theta$ .*

### 1.1.5 The extension problem.

Let  $(\Omega, \mathcal{C}, P)$  be a finitely additive probability space. The (countably additive) probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be an *extension* of  $(\Omega, \mathcal{C}, P)$  if  $\mathcal{C} \subset \mathcal{F}$  and  $\mathbb{P}(A) = P(A)$  for every  $A \in \mathcal{C}$ . Speaking more loosely, we also say that  $\mathbb{P}$  extends  $P$  from  $\mathcal{C}$  to  $\mathcal{F}$ . Typically, one builds a probability space by first defining a finitely additive probability space on an algebra, and then proving it has an extension. As we saw in the previous section, algebras are easy to construct and thus explicitly defining finitely additive probability spaces is generally straightforward. Once this is done, the following result from measure theory can be used to decide if there is a countably additive extension or not.

**Theorem 6** Carathéodory’s Extension Theorem *Let  $P$  be a finitely-additive probability measure on an algebra  $\mathcal{C}$ . Then  $P$  admits an extension to a probability measure  $\mathbb{P}$  on  $\sigma(\mathcal{C})$  if and only if  $P$  is continuous from above at  $\emptyset$ . The extension  $\mathbb{P}$  is unique.*

We will not give the full proof of this theorem. It can be found in textbooks on measure and integration theory. For example, the book of Folland cited above presents a concise, but complete proof in Chapter 1. We recall only the main steps. First one defines the outer measure  $P^*$  associated to  $P$ :

$$\text{for every } E \subset \Omega, \text{ let } P^*(A) = \inf \left\{ \sum_1^\infty P(U_i); E \subset \bigcup_1^\infty U_i, \{U_i\} \subset \mathcal{C} \right\}.$$

Next one defines  $A \subset \Omega$  to be  $P^*$ -measurable if

$$P^*(E) = P^*(A \cap E) + P^*(E \cap A^c) \quad \text{for all } E \subset \Omega.$$

Then one proves that if  $P$  is continuous from above at  $\emptyset$ ; (i) the collection of  $P^*$  measurable sets is a  $\sigma$ -algebra including  $\mathcal{C}$ ; (ii) restricted to the  $P^*$ -measurable subsets,  $P^*$  is a countably additive measure; and (iii)  $P^*(A) = P(A)$  if  $A \in \mathcal{C}$ . Let  $\mathcal{M}$  denote the  $P^*$ -measurable sets. By (i),  $\sigma(\mathcal{C}) \subset \mathcal{M}$ . Let  $\tilde{\mathbb{P}}$  denote  $P^*$  restricted to  $\mathcal{M}$ , and let  $\mathbb{P}$  denote  $P^*$  restricted to  $\sigma(\mathcal{C})$ . It follows that  $(\Omega, \sigma(\mathcal{C}), \mathbb{P})$  is an extension of  $(\Omega, \mathcal{C}, P)$ , and  $(\Omega, \mathcal{M}, \tilde{\mathbb{P}})$  is an extension of  $(\Omega, \sigma(\mathcal{C}), \mathbb{P})$ . It can also be shown that  $(\Omega, \mathcal{M}, \tilde{\mathbb{P}})$  is *complete*; this means that if  $A \subset B$ , where  $B \in \mathcal{M}$  and  $\tilde{\mathbb{P}}(B) = 0$ , then it follows that  $A \in \mathcal{M}$ . (In fact,  $(\Omega, \mathcal{M}, \tilde{\mathbb{P}})$  is the completion of  $(\Omega, \sigma(\mathcal{C}), \mathbb{P})$ ; see Folland, Chapter 1.)

The proof just outlined yields an extension of  $P$  to a  $\sigma$ -algebra  $\mathcal{M}$  that may be larger than  $\sigma(\mathcal{C})$ , so we could have stated a stronger theorem. However, for this course, we will mostly just use the extension to  $\sigma(\mathcal{C})$ , and this will suffice for our purposes. One reason for restricting to  $\sigma(\mathcal{C})$  is the following. Suppose one is potentially interested in constructing lots of other probability measures, starting from finitely additive measures on  $(\Omega, \mathcal{C})$ . The  $\sigma$ -algebra of  $P^*$ -measurable sets depends on  $P$ , but  $\sigma(\mathcal{C})$  does not. Thus  $\sigma(\mathcal{C})$  will provide a universal class of events for a large family of probability spaces. (Sometimes, on the other hand, it is important for technical reasons to have a complete probability space and then one will use the extension  $(\Omega, \mathcal{M}, \tilde{\mathbb{P}})$ .)

Examples of the use of Carathéodory's extension theorem are given in the next section.

## 1.2 Examples of Probability Spaces

### 1.2.1 Finite tosses of a fair coin.

Our object is to construct a probability space to model  $N$  flips of a fair coin. If each occurrence of a head is represented by the number 1 and each occurrence of a tail by the number 0, the outcome of a sequence of  $N$  tosses is an  $N$ -vector of 0's and 1's. Therefore the outcome space can be represented by  $\Omega_N = \{0, 1\}^N$ . This is a finite set and for the  $\sigma$ -algebra of events we shall use its power set.

If the coin is fair, all sequences should have equally likely outcomes; hence, the appropriate probability measure is the uniform measure defined in Example 1.4. Since the cardinality of  $\{0, 1\}^N$  is  $2^N$ ,

$$\mathbb{P}_N(B) \triangleq \frac{|B|}{2^N}, \quad \text{for } B \subset \{0, 1\}^N.$$

The probability of any specific sequence of heads and tails is thus  $1/2^N$ .

From the point of view of theory, we are done. We have an explicit formula for the probability of any event. Of course, if  $N$  is large and someone hands you a particular  $B$ , actually finding  $|B|$  may be a hard combinatorial problem. This is really where the fun starts.

An easy problem from elementary probability is to find the probability that there are exactly  $m$  heads in  $N$  tosses. This is represented by the event  $A$  consisting of all sequences in  $\{0, 1\}^N$  with exactly  $m$  1's. The cardinality of this set is the number of ways to choose  $m$  from  $N$ . Hence

$$\mathbb{P}_N(A) = \binom{N}{m} \frac{1}{2^N}.$$

### 1.2.2 Infinite tosses of a fair coin.

A major theme of probability is the statistical behavior of the outcomes of random trial as the number of repetitions tends to infinity. For example, one would like to know under what circumstances the limiting empirical frequency of an event does indeed converge to its probability, as the frequency interpretation of probability requires. To study these issues for coin tossing, it is convenient to construct a probability space that supports an infinite number of trials.

As a first step, we define a finitely additive probability space. Generalizing from the finite case, we use for  $\Omega$  the infinite product space,  $\Omega = \{0, 1\}^\infty$ , consisting of all countable sequences  $\omega = (\omega_1, \omega_2, \dots)$  of 0's and 1's.

The algebra of events should include at least all those events defined in terms of what happens to a finite number of tosses only. For every integer  $N \geq 1$ , let  $\pi_N : \{0, 1\}^\infty \rightarrow \{0, 1\}^N$  be the projection of  $\{0, 1\}^\infty$  onto its first  $N$  coordinates:  $\pi_N((\omega_1, \omega_2, \dots)) = (\omega_1, \dots, \omega_N)$ . Then any event in  $\{0, 1\}^\infty$  that depends only on the first  $N$  tosses takes the form  $\pi_N^{-1}(B) = \{\omega \in \{0, 1\}^\infty; (\omega_1, \dots, \omega_N) \in B\}$ . The family of such events,

$$\mathcal{C}_N \triangleq \{\pi_N^{-1}(B); B \subset \{0, 1\}^N\}.$$

is a finite algebra (and hence a  $\sigma$ -algebra). The union,

$$\mathcal{C} \triangleq \bigcup_1^\infty \mathcal{C}_n.$$

is the collection of all events determined by the outcome of a finite number of tosses. It is easy to verify it is an algebra (exercise!), and its elements are called *cylinder set* or *cylinder events*. It is not a  $\sigma$ -algebra. For example, if

$\omega^* = (\omega_{*1}, \omega_{*2}, \dots)$  is a given point in  $\{0, 1\}^\infty$ , then the singleton set  $\{\omega^*\}$  is not in  $\mathcal{C}$ . But

$$\{\omega^*\} = \bigcap_{n=1}^{\infty} \{\omega; \omega_n = \omega_n^*\},$$

so  $\{\omega^*\}$  is a countable intersection of sets in  $\mathcal{C}$ .

Finally, we define a finitely additive probability measure on  $\mathcal{C}$ . This measure should give results consistent with the previous model for a finite number of tosses of a fair coin. Therefore, if  $B \subset \{0, 1\}^N$ , we want

$$P\left(\pi_N^{-1}(B)\right) = \mathbb{P}_N(B) = \frac{|B|}{2^N} \quad (1.6)$$

In order to show that this definition is consistent we must show that if  $B_1 \in \{0, 1\}^{N_1}$ ,  $B_2 \in \{0, 1\}^{N_2}$ , and  $\pi_{N_1}^{-1}(B_1) = \pi_{N_2}^{-1}(B_2)$ , then

$$\frac{|B_1|}{2^{N_1}} = \frac{|B_2|}{2^{N_2}}.$$

It is not hard to show that if  $N_1 = N_2$ , then  $\pi_{N_1}^{-1}(B_1) = \pi_{N_2}^{-1}(B_2)$  if and only if  $B_1 = B_2$ , so there is no problem in this case. Suppose, without loss of generality, that  $N_1 < N_2$ . Then  $\pi_{N_1}^{-1}(B_1) = \pi_{N_2}^{-1}(B_2)$  if and only if  $B_2 = B_1 \times \{0, 1\}^{N_2 - N_1}$ . Thus

$$\frac{|B_2|}{2^{N_2}} = \frac{|B_1|2^{N_2 - N_1}}{2^{N_2}} = \frac{|B_1|}{2^{N_1}}.$$

The measure  $P$  is certainly finitely additive on  $\mathcal{C}$ . To see this, let  $U, V \in \mathcal{C}$  and assume they are disjoint. Since  $U$  and  $V$  each are events concerned with a finite number of tosses, there is an integer  $N \geq 1$  and subsets  $A, B$  of  $\{0, 1\}^N$  such that  $U = \pi_N^{-1}(A)$  and  $V = \pi_N^{-1}(B)$ . Because  $U$  and  $V$  are disjoint, so also are  $A$  and  $B$ . Thus

$$P(U \cup V) = P\left(\pi_N^{-1}(A \cup B)\right) = \frac{|A \cup B|}{2^N} = \frac{|A| + |B|}{2^N} = P(U) + P(V).$$

The triple we have constructed,  $(\{0, 1\}^\infty, \mathcal{C}, P)$ , is a finitely additive probability space for an infinite number of coin tosses. Next, we want to use Carathéodory's theorem to define a countably additive extension. Actually, we will do more. We will show that any finitely additive probability measure on  $(\{0, 1\}^\infty, \mathcal{C})$  extends to a countably additive probability measure on  $(\{0, 1\}^\infty, \sigma(\mathcal{C}))$ . Using this result (Theorem 8 below) it follows that there is

a probability space  $(\{0, 1\}^\infty, \sigma(\mathcal{C}), \mathbb{P})$  extending the finitely additive coin toss model  $(\{0, 1\}^\infty, \mathcal{C}, P)$ .

The crucial ingredient of the proof of the extension is a result having nothing to do with measure theory. For each positive integer  $i \geq 1$ , assume  $\Theta_i$  is a non-empty finite set and consider the product space  $\Theta = \Theta_1 \times \Theta_2 \times \cdots = \bigotimes_{i=1}^\infty \Theta_i$ . Let  $\pi_n$  again denote projection of  $\Theta$  on its first  $N$  coordinates, and again let  $\mathcal{C}$ , the algebra of cylinder sets, be the collection of all subsets of the form  $\pi_N^{-1}(B)$  for some  $N \geq 1$  and  $B \subset \bigotimes_{i=1}^N \Theta_i$ .

**Theorem 7** *Let  $\{A_n\}$  be a decreasing sequence of subsets of  $\Theta$  such that  $A_n \in \mathcal{C}$  for every  $n$  and  $\bigcap_1^\infty A_n = \emptyset$ . Then there exists an  $N$  such that  $A_n = \emptyset$  for all  $n \geq N$ .*

**Proof I:** Here is a short proof for those who know Tychonoff's theorem, which says that a product of compact spaces is compact in the product topology. A direct proof will be presented later. If each  $\Theta_i$  is supplied with the discrete topology, in which every subset is open and closed, each  $\Theta_i$  is compact, and hence  $\Theta$  is compact in the product topology. But every cylinder set is closed and hence compact. A decreasing sequence of non-empty compact sets has a nonempty intersection. Therefore, if the intersection of a decreasing sequence of cylinder sets has an empty intersection, the sets in the sequence must be empty from some finite  $N$  on.  $\diamond$

**Theorem 8** *Any finitely additive probability measure  $P$  on  $(\Theta, \mathcal{C})$  extends to countably additive measure on  $(\Theta, \sigma(\mathcal{C}))$ .*

**Proof:** If  $\{A_n\}$  is a sequence of cylinder sets decreasing to  $\emptyset$ , then  $A_n = \emptyset$  for  $n$  sufficiently large. Hence  $P(A_n) = 0$  for  $n$  sufficiently large, implying  $\lim_{n \rightarrow \infty} P(A_n) = 0$ . Thus  $P$  is continuous from above at  $\emptyset$  and hence, by Carathéodory's theorem, it admits an extension.  $\diamond$

**Proof II of Theorem 7:** This is a direct proof. We will prove the equivalent statement: if  $\{A_n\}$  is a decreasing sequence of cylinder sets and  $A_n \neq \emptyset$  for every  $n$ , then  $\bigcap_1^\infty A_n \neq \emptyset$ .

Let  $A \subset \Theta$ . For any finite sequence  $(\xi_1, \dots, \xi_N) \in \bigotimes_1^N \Theta_i$ , the subset

$$A(\xi_1, \dots, \xi_N) = \{(\omega_{N+1}, \omega_{N+2}, \dots) \in \times_{N+1}^\infty \Theta_i; (\xi_1, \dots, \xi_N, \omega_{N+1}, \omega_{N+2}, \dots) \in A\}$$

of  $\bigotimes_{i=N+1}^\infty \Theta_i$  is the *section* of  $A$  at  $(\xi_1, \dots, \xi_N)$ .

**Lemma 3** *If  $\{A_n\}$  is a decreasing sequence of subsets of  $\Theta$  such that  $A_n$  is nonempty for every  $n$ , there exists a  $\omega^* = (\omega_1^*, \omega_2^*, \dots)$  such that  $A_n(\omega_1^*, \dots, \omega_N^*)$  is nonempty for every  $n$  and every  $N$ .*

This lemma is proved inductively on  $N$ . The case  $N = 0$  is just the statement that  $A_N$  is nonempty for all  $N$ , which is true by assumption. In order to simplify the notation, we shall carry out in detail only the induction step going from  $N = 0$  to  $N = 1$ .

The crucial observation is that

$$A_n = \bigcup_{\xi_1 \in \Theta_1} \{\xi_1\} \times A_n(\xi_1). \quad (1.7)$$

Notice that for each  $\xi_1$ ,  $\{A_n(\xi_1)\}$  is a decreasing sequence of subsets of  $\bigotimes_{i=2}^{\infty} \Theta_i$ . Thus if for some  $M$ ,  $A_M(\xi_1) = \emptyset$ ,  $A_n(\xi_1) = \emptyset$  for all  $n \geq M$ . Since  $\Theta_1$  is finite and since  $A_n$  is nonempty for all  $n$ , it follows from (1.7) that there must exist at least one  $\omega_1 \in \Theta_1$  such that  $A_n(\omega_1)$  is nonempty for all  $n$ . Take  $\omega_1^*$  to be such an  $\omega_1$ .

Now suppose we have found  $(\omega_1^*, \dots, \omega_k^*)$  such that  $A_n(\omega_1^*, \dots, \omega_k^*)$  is nonempty for all  $n$ . Then by applying the same argument we just made with  $A_n(\omega_1^*, \dots, \omega_k^*)$  in place of  $A_n$ , we find there exists  $\omega_{k+1}^*$  such that  $A_n(\omega_1^*, \dots, \omega_k^*, \omega_{k+1}^*)$  is nonempty for all  $n$ . This completes the induction step.  $\diamond$

Let  $\omega^* = (\omega_1^*, \omega_2^*, \dots)$  be the point in  $\Theta$  provided by the Lemma for a decreasing sequence of nonempty cylinder sets. We will show that  $\omega^* \in A_n$  for all  $n$  and hence that  $\omega^* \in \bigcap A_n$ . Let  $n$  be arbitrary. Since  $A_n$  is a cylinder set, there is an  $N$  and a  $B \subset \bigotimes_{i=1}^N \Theta_i$  such that  $A_n = \pi_N^{-1}(B)$ . Since  $A_n(\omega_1^*, \dots, \omega_N^*)$  is not empty,  $(\omega_1^*, \dots, \omega_N^*) \in B$ . But then  $\pi_N(\omega^*) \in B$ , which implies  $\omega^* \in A_n$ . This completes the proof of Theorem 7.  $\diamond$

### 1.2.3 Probability Measures on $\mathbb{R}$ .

Recall that  $\mathcal{B}(\mathbb{R})$  denotes the collection of Borel sets of  $\mathbb{R}$ . This section discusses how to construct probability measures on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ .

First suppose we are given a probability measure  $\mathbb{P}$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Define the function  $F : \mathbb{R} \rightarrow [0, 1]$ ,

$$F_{\mathbf{P}}(x) = \mathbb{P}\left((-\infty, x]\right), \quad x \in \mathbb{R}. \quad (1.8)$$

**Lemma 4**  *$F_{\mathbf{P}}$  satisfies: (i)  $F$  is nondecreasing and right-continuous; (ii)  $\lim_{x \rightarrow -\infty} F_{\mathbf{P}}(x) = 0$ ; (iii)  $\lim_{x \rightarrow \infty} F_{\mathbf{P}}(x) = 1$ .*

**Proof:** If  $y > x$ ,  $(-\infty, x] \subset (-\infty, y]$  and hence  $F_{\mathbf{P}}(x) \leq F_{\mathbf{P}}(y)$ , proving  $F_{\mathbf{P}}$  is nondecreasing. All other properties are consequences of the continuity of  $\mathbb{P}$ , in the measure-theoretic sense. For example, since  $(-\infty, x] = \bigcap_{n=1}^{\infty} (-\infty, x + (1/n)]$ , it follows from continuity from above of  $\mathbb{P}$ , that  $\lim_{n \rightarrow \infty} F_{\mathbf{P}}(x + (1/n)) = F_{\mathbf{P}}(x)$ . Together with the fact that  $F$  is nondecreasing, this proves  $F_{\mathbf{P}}$  is right continuous. Property (ii) is true by continuity of  $\mathbb{P}$  from above at  $\emptyset$ , because  $\emptyset = \bigcap_{n=1}^{\infty} (-\infty, -n]$ . Likewise, since  $\bigcup_{n=1}^{\infty} (-\infty, n] = (-\infty, \infty)$  and  $\mathbb{P}((-\infty, \infty)) = 1$ , (iii) is a consequence of the continuity from below of  $\mathbb{P}$ .  $\diamond$

**Definition 4** *A function  $F$  satisfying conditions (i)–(iii) of Lemma 4 is called a probability distribution function.*

If one is trying to construct a  $\mathbb{P}$  to model a random trial whose outcome is a real number, it is natural to start by prescribing the probabilities of intervals, which are the simplest events. This is equivalent to prescribing  $F_{\mathbf{P}}$ . Therefore, one would like construct probability measures starting from probability distribution functions.

**Theorem 9** *Let  $F$  be a probability distribution function. There is a unique probability measure  $\mathbb{P}_F$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  such that  $F(x) = \mathbb{P}_F((-\infty, x])$ , for  $x \in \mathbb{R}$ .*

**Proof:** Let  $\mathcal{E}$  be the collection of all intervals that take one of the forms  $(a, b]$ ,  $(-\infty, b]$ ,  $(a, \infty)$ , or  $(-\infty, \infty)$ . Let  $\mathcal{R}$  be the class of finite disjoint unions of intervals in  $\mathcal{E}$ . We know that  $\mathcal{R}$  is an algebra (see Example 1.5) and it is easy to see that  $\sigma(\mathcal{R}) = \mathcal{B}(\mathbb{R})$  (exercise).

Define

$$P((a, b]) = F(b) - F(a), \quad P((-\infty, b]) = F(b), \quad P((a, \infty)) = 1 - F(a), \quad P(\mathbb{R}) = 1.$$

and extend  $P$  to  $\mathcal{R}$  by finite additivity. To check that this definition is consistent we must show, for instance, that if an interval  $I = (a, b]$  in  $\mathcal{E}$  is represented as a finite disjoint union,  $I = \bigcup_1^m I_i$  of intervals in  $\mathcal{E}$ , then  $F(b) - F(a) = \sum_1^m P(I_i)$ ; this is left as an exercise. We now have a finitely additive probability measure  $P$  on  $(\mathbb{R}, \mathcal{R})$ .

We will show that  $P$  extends to a unique probability measure on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  by using Carathéodory's extension theorem. As in the proof of Theorem 8, the argument exploits compactness.

We first claim that if  $I \in \mathcal{E}$ , then, for any  $\epsilon > 0$ , there is an  $I_0 \in \mathcal{E}$  such that the closure  $\bar{I}_0$  is compact and contained in  $I$  and

$$\epsilon > P(I - I_0) = P(I) - P(I_0) > 0.$$

Indeed, let  $I = (a, b]$  and  $I_0 = (a_0, b]$ , where  $a < a_0 < b$ . Then  $P(I - I_0) = P((a, a_0]) = F(a_0) - F(a)$ . Since  $F$  is right-continuous, given any  $\epsilon > 0$ , there is an  $a_0$ , with  $a < a_0 < b$  such that  $F(a) - F(a_0) < \epsilon$ . Since  $\bar{I}_0 = [a_0, b] \subset (a, b]$ , we have proved the claim for  $I = (a, b]$ . The other cases ( $I = (-\infty, b]$ , or  $(a, \infty)$  or  $\mathbb{R}$ ) are proved in a similar fashion.

Now, since any element of  $\mathcal{R}$  is a finite disjoint union of intervals of  $\mathcal{E}$ , it easily follows that if  $A \in \mathcal{R}$  then for any  $\epsilon > 0$ , there is an  $A_0 \in \mathcal{R}$  such that  $\bar{A}_0 \subset A$ ,  $\bar{A}_0$  is compact, and  $P(A - A_0) < \epsilon$ .

Let  $\{A_n\}$  be a decreasing sequence of sets in  $\mathcal{R}$  such that  $\bigcap_{n=1}^{\infty} A_n = \emptyset$ . Let  $\epsilon > 0$  and for each take  $B_n \in \mathcal{R}$  such that  $\bar{B}_n$  is compact,  $\bar{B}_n \subset A_n$ , and  $P(A_n - B_n) < \epsilon/2^n$ . Let  $C_n = \bigcap_{i=1}^n B_i$ . Observe first that

$$A_n - C_n = \bigcup_{i=1}^n (A_n - B_i) \subset \bigcup_{i=1}^n (A_i - B_i).$$

Therefore,

$$P(A_n - C_n) \leq \sum_{i=1}^n P(A_i - B_i) \leq \sum_{i=1}^n \frac{\epsilon}{2^i} < \epsilon \quad \text{for every } n \geq 1.$$

Now  $\{\bar{C}_n\}$  is a decreasing sequence of compact sets and  $\bigcap_{n=1}^{\infty} \bar{C}_n \subset \bigcap_{n=1}^{\infty} A_n = \emptyset$ . Therefore, there exists a positive integer  $N$  such that  $\bar{C}_n = \emptyset$ , and thus  $C_n = \emptyset$ , for  $n \geq N$ . Hence it follows that for  $n \geq N$ ,  $P(A_n) \leq P(C_n) + \epsilon = \epsilon$ . Therefore,

$$\limsup_{n \rightarrow \infty} P(A_n) \leq \epsilon.$$

But this is true for any  $\epsilon > 0$  and so taking  $\epsilon \downarrow 0$ , we find that

$$\lim_{n \rightarrow \infty} P(A_n) = 0,$$

thereby proving that  $P$  on  $(\mathbb{R}, \mathcal{R})$  is continuous from above at  $\emptyset$ . Carathéodory's extension theorem says then says that there is a unique probability measure on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  extending  $P$ .

If  $\mathbb{P}'$  is a probability measure on the Borel sets such that  $\mathbb{P}'((-\infty, x]) = F(x)$  for all  $x$ , then for any  $a < b$ ,  $\mathbb{P}'((a, b]) = \mathbb{P}'((-\infty, b]) - \mathbb{P}'((-\infty, a]) = F(b) -$

$F(a)$ . Thus  $\mathbb{P}'$  must agree with  $P$  on  $\mathcal{R}$  and hence be an extension of  $P$ . But by Carathéodory's theorem, the extension is unique and so  $\mathbb{P}' = \mathbb{P}$ .  $\diamond$

Students of real analysis will recognize that in Theorem 9 we have constructed the Lebesgue-Stieltjes measure associated to  $F$ .

Theorem 9 has a generalization to  $\mathbb{R}^n$  that we will discuss in the chapter on random variables.

**Example 1.6.** Let

$$F(x) = \begin{cases} 0, & \text{if } x \leq 0; \\ x, & \text{if } 0 < x \leq 1; \\ 1, & \text{if } x \geq 1. \end{cases}$$

Let  $\mathbb{P}$  be the measure  $\mathbb{P}_F$  restricted to the Borel sets contained in  $[0, 1]$ . Then  $\mathbb{P}$  is the countably additive extension of the finitely additive uniform measure introduced in Example 1.2. It is Lebesgue measure restricted to  $[0, 1]$ .

## 1.3 Independence

### 1.3.1 Definitions and simple examples.

The concept of independence is central to probability theory. It formalizes the notion of different random events that do not affect each others outcomes statistically.

**Definition 5** *Two events  $A$  and  $B$  in a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  are independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.9}$$

This definition is best understood using the concept of conditional probability. Suppose  $U$  and  $V$  are two events and  $\mathbb{P}(V) > 0$ . The ratio

$$\mathbb{P}(U|V) = \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(V)}$$

is called the conditional probability of  $U$  given  $V$ . It is interpreted as the probability  $U$  occurs given one knows  $V$  has occurred, but nothing else.

If  $A$  and  $B$  satisfy condition (1.9) and  $\mathbb{P}(B) > 0$ , then that  $\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B)/\mathbb{P}(B) = \mathbb{P}(A)$ . Thus knowledge that  $B$  has occurred does not affect at all our assessment of the probability of  $A$ . This explains the sense in which  $A$  and  $B$  are independent. The definition is stated without reference to conditional

probabilities because the product condition (1.9) is symmetric in  $A$  and  $B$  and it covers the case in which  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(B) = 0$  and conditioning on  $A$  (respectively  $B$ ) is not defined.

Notice that if  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(A) = 1$ , then  $A$  is independent of all other events. Since  $A \cap A = A$ , an event is independent of itself if and only if  $\mathbb{P}(A) = \mathbb{P}^2(A)$ , which is true if and only if  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(A) = 1$ .

Let  $\Omega$  be an outcome space and  $\mathcal{F}$  a  $\sigma$ -algebra of events. Let  $\mathbb{P}_0$  and  $\mathbb{P}_1$  be two, different probability measures on  $(\Omega, \mathcal{F})$ . Of course, it may well be that events  $A$  and  $B$  in  $\mathcal{F}$  are independent when  $\mathbb{P}_0$  is the probability measure, but not when  $\mathbb{P}_1$  is the probability measure. When clarity is needed, we will use the locution, “ $A$  and  $B$  are independent with respect to  $\mathbb{P}$ ,” to indicate the probability measure being used. When we just say, “ $A$  and  $B$  are independent,” we have in mind a fixed probability space on which we are working.

**Example 1.7.** Consider the probability space, constructed in the previous section, modeling repeated tosses of a fair coin. Let  $A$  be the event that toss  $i$  is heads and  $B$  be the event toss  $j$  is heads, where  $i \neq j$ . Let  $N = \max i, j$ . Since any sequence of  $N$  tosses is equally likely, then by reordering the sequence it is clear that we may assume  $i = 1$  and  $j = 2$ . Then, using the formula and notation of equation (1.6),  $\mathbb{P}(A \cap B) = \mathbb{P}_2(\{(1, 1)\}) = \frac{1}{2^2} = \frac{1}{4}$ . At the same time  $\mathbb{P}(A) = \mathbb{P}_1(\{1\}) = 1/2$  and similarly,  $\mathbb{P}(B) = 1/2$ . Therefore  $A$  and  $B$  are independent.

By repeating the argument used in Example 1.7, one can show that any event concerning the outcome of toss  $i$  is independent of any event concerning the outcome of toss  $j$ , if  $i \neq j$ . Many situations like this occur and it is convenient to define independence between families of events.

**Definition 6** *Let  $\mathcal{A}$  and  $\mathcal{B}$  be families of events in a probability space. Then  $\mathcal{A}$  and  $\mathcal{B}$  are independent if  $A$  and  $B$  are independent for every  $A \in \mathcal{A}$  and  $B \in \mathcal{B}$ .*

The next lemma is a simple application of this concept. It illustrates a point to which we return repeatedly in these notes: in probability theory, we often concern ourselves not with the analysis of a single event, but with a collection of events that is a sub- $\sigma$ -algebra of the collection of all events. For example, if  $A$  is an event, the smallest  $\sigma$ -algebra of events containing  $A$  is the finite collection,  $\{A, A^c, \Omega, \emptyset\}$ . It is natural to consider this whenever thinking about  $A$ , since we are implicitly thinking about whether  $A$  occurs or about whether  $A^c$  occurs.

**Lemma 5** *Events  $A$  and  $B$  in a probability space are independent if and only if  $\{A, A^c, \Omega, \emptyset\}$  is independent of  $\{B, B^c, \Omega, \emptyset\}$ .*

**Proof:** The “if” direction is trivial. For the “only if” direction, it suffices to check that  $A$  and  $B^c$  are independent, knowing that  $A$  and  $B$  are independent. Indeed,  $\Omega$  and  $\emptyset$  are automatically independent of all other events, and, once we have determined the independence of  $A$  and  $B^c$ , the other cases follow by first switching the roles of  $A$  and  $B$  and then the roles of  $A$  and  $A^c$ .

For the independence of  $A$  and  $B^c$ , observe that  $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A)\mathbb{P}(B) + \mathbb{P}(A \cap B^c)$ . From this, it follows that  $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)[1 - \mathbb{P}(B)] = \mathbb{P}(A)\mathbb{P}(B^c)$ , as we wanted to prove.  $\diamond$

**Example 1.8.** In the fair coin example, let  $\mathcal{A}_i$  be the  $\sigma$ -algebra generated by all events concerning toss  $i$ ; this is simply  $\{A_i, A_i^c, \Omega, \emptyset\}$ , where  $A_i$  is the probability of heads on toss  $i$ . Then it is a consequence of Example 1.7 and Lemma 5 that  $\mathcal{A}_i$  and  $\mathcal{A}_j$  are independent of one another. (Actually, using Lemma 5 to prove this obscures what is really going on in the coin toss example, which is that the probability measure is a product measure. We return to this point below.)

From Examples 1.7 and 1.8 any two coin in the fair toss model are independent of one another. We could go on to show that any toss  $i$  is independent of any event concerning two, three, or any number of other tosses. There is more going on in this example than just pairwise independence of two different tosses. The next definition formalizes the meaning of mutual independence between more than one event.

**Definition 7** *Events  $A_1, \dots, A_n$  are (mutually) independent if for every  $2 \leq r \leq n$  and  $1 \leq i_1 < i_2 < \dots < i_r$ ,*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_r}) \quad (1.10)$$

*Let  $\mathcal{E}$  be a family of events. The events of  $\mathcal{E}$  are (mutually) independent if the events of every finite subset of  $\mathcal{E}$  are mutually independent.*

*Let  $\{\mathcal{A}_\alpha \in \mathcal{I}\}$  be an arbitrary collection of families of events. The families of this collection are (mutually) independent if for any finite subset of distinct indices  $\{\alpha_1, \dots, \alpha_n\}$  from  $\mathcal{I}$ , and any events  $A_1 \in \mathcal{A}_{\alpha_1}, A_2 \in \mathcal{A}_{\alpha_2}, \dots, A_n \in \mathcal{A}_{\alpha_n}$ ,  $A_1, A_2, \dots, A_n$  are independent.*

To check, for example, that  $A_1, A_2, A_3$  are (mutually) independent, it is necessary to check that they are pairwise independent ( $r = 2$  in the definition above), and, in addition, that  $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$ .

In practice, checking that the families of the collection  $\{\mathcal{A}_\alpha \in \mathcal{I}\}$  are independent is a bit simpler than direct applications of the definitions above would indicate. It is only necessary to show that for any finite subset of distinct indices  $\{\alpha_1, \dots, \alpha_n\}$  from  $\mathcal{I}$ , and any events  $A_1 \in \mathcal{A}_{\alpha_1}, A_2 \in \mathcal{A}_{\alpha_2}, \dots, A_n \in \mathcal{A}_{\alpha_n}$ ,

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n).$$

This identity by itself does not prove that  $A_1, \dots, A_n$  are independent. But if we show this identity for all possible finite subsets of indices and associated choices of events, we will have verified (1.10) for  $A_1, \dots, A_n$ , thus proving their independence.

If  $A_1, \dots, A_n$  are mutually independent, then they are of course pairwise independent. However, a collection of events may be pairwise independent without being mutually independent. The reader should create an example as an exercise. The word “mutually” was used here to emphasize the distinction between Definition 7 and pairwise independence. But as mutual independence is the usual case, we hereafter use “independent” for “mutually independent.” When a collection of events is only pairwise independent, we will say so explicitly.

**Example 1.9.** We continue with the coin tossing example. Let  $\mathcal{A}_i, i \geq 1$  be defined as in Example 1.8. Then the families  $\mathcal{A}_1, \mathcal{A}_2, \dots$  are independent. To show this let  $1 \leq j_1 < j_2 < \dots < j_r = N$ , and let  $U_{j_k} \in \mathcal{A}_{j_k}$  for each  $1 \leq k \leq r$ . We can write

$$\bigcap_{k=1}^r U_{j_k} = \bigcap_{i=1}^N U_i, \quad (1.11)$$

where, if  $i \notin \{j_1, \dots, j_k\}$ , then  $U_i = \Omega$ . Then, we will be done if we can show,

$$\mathbb{P}\left(\bigcap_{i=1}^N U_i\right) = \prod_{i=1}^N \mathbb{P}(U_i). \quad (1.12)$$

Since  $\mathbb{P}(U_i) = 1$  for  $i \notin \{j_1, \dots, j_k\}$ , equations (1.11) and (1.12) together will imply

$$\mathbb{P}\left(\bigcap_{k=1}^r U_{j_k}\right) = \prod_{i=1}^r \mathbb{P}(U_{j_i}),$$

and by the remark after Definition 7, this is what we need to show for arbitrary choices of indices  $j$  and events in  $\mathcal{A}_j$  in order to prove independence. Therefore we concentrate on showing (1.4) for an arbitrary  $N$  and arbitrary choice of  $U_i \in \mathcal{A}_i$ , for  $1 \leq i \leq N$ .

Since each  $U_i$  is an event concerning the outcome of toss  $i$ , each  $U_i$  can be written in the form  $U_i = \{(\omega_1, \omega_2, \dots); \omega_i \in B_i\}$ , where  $B_i$  is a subset of  $\{0, 1\}$ . Moreover,

$$\mathbb{P}(U_i) = \frac{|B_i|}{2} \quad (1.13)$$

Now,

$$\bigcap_{i=1}^N U_i = \{(\omega_1, \omega_2, \dots); \omega_i \in B_i, 1 \leq i \leq n\} = B_1 \times \dots \times B_N \times \{0, 1\} \times \{0, 1\} \times \dots.$$

This is a cylinder set and its probability is

$$\mathbb{P}\left(\bigcap_{i=1}^N U_i\right) = \frac{|B_1 \times \dots \times B_N|}{2^N} = \prod_{i=1}^N \frac{|B_i|}{2}.$$

Because of equation (1.13), this proves (1.12).  $\diamond$

**Example 1.10.** Let  $\mathbb{P}$  be the uniform probability measure on the Borel sets of  $[0, 1]$ , as constructed in Example 1.6. Any  $x \in [0, 1]$ , has a unique representation

$$x = \sum_{i=0}^{\infty} \frac{x_i}{2^i},$$

where  $\{x_i\}$  is a sequence of 0's and 1's that does not end in a nonterminating string of 1's. For  $i \geq 1$ , let  $A_i = \{x \in [0, 1]; x_i = 1\}$  and let  $\mathcal{A}_i = \{A_i, A_i^c, \Omega, \emptyset\}$ . Then  $\mathcal{A}_1, \mathcal{A}_2, \dots$  are independent. Proving this is left as an exercise. The events in  $\mathcal{A}_1, \mathcal{A}_2, \dots$  provide an alternate model of independent tosses of a fair coin, as we shall see in Chapter 2 on random variables.

The following lemma is sometimes useful for checking independence of a collection of families of events.

**Lemma 6** *Let  $\mathcal{A}_1, \dots, \mathcal{A}_N$  be families of events and assume that  $\Omega \in \mathcal{A}_i$  for each  $i$ . Then the families are independent of one another if*

$$\mathbb{P}(A_1 \cap \dots \cap A_N) = \mathbb{P}(A_1) \dots \mathbb{P}(A_N) \quad (1.14)$$

whenever  $A_1 \in \mathcal{A}_1, \dots, A_N \in \mathcal{A}_N$ .

**Proof:** The point here is that to check independence, we must show that for any  $r \leq n$ , any  $1 \leq i_1 < \dots < i_r \leq n$ , and any  $A_{i_j} \in \mathcal{A}_{i_j}$ ,  $\mathbb{P}\left(\bigcap_{j=1}^r A_{i_j}\right) = \prod_{j=1}^r \mathbb{P}(A_{i_j})$ . But these cases are all included in (1.2) assuming  $\Omega \in \mathcal{A}_i$  for each  $i$ . It is clearer to illustrate by example, then develop the messy notation for a general proof. For example, the following calculation shows that any event from  $\mathcal{A}_1$  is independent of any event from  $\mathcal{A}_2$ :

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_2 \cap \Omega \cap \dots \cap \Omega) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(\Omega) \cdots \mathbb{P}(\Omega) = \mathbb{P}(A_1)\mathbb{P}(A_2).$$

### 1.3.2 Tossing a loaded coin.

The object in this section is to construct probability space for repeated independent tosses of a coin for which the probability of heads is  $p$  and the probability of tails is  $q = 1 - p$ . We use the notation developed for the fair coin toss spaces. For a single coin toss the probability space is  $\{0, 1\}$  and we define the probability measure by  $\mathbb{P}_1(1) = p$  and  $\mathbb{P}(\{0\}) = q$ . It is convenient to note that for  $x \in \{0, 1\}$ , we can write this as  $\mathbb{P}_1(\{x\}) = p^x q^{1-x}$ .

To define the probability measure on  $\{0, 1\}^N$  for  $N$  independent tosses, we need only define the probability of every singleton outcome. But the assumption that the outcomes of different tosses are independent requires,

$$\begin{aligned} \mathbb{P}_N(\{(\omega_1, \dots, \omega_N)\}) &= \prod_{i=1}^N (\text{probability toss } i \text{ results in } \omega_i) \\ &= \prod_{i=1}^N p^{\omega_i} q^{1-\omega_i} = p^{\sum_{i=1}^N \omega_i} q^{N - \sum_{i=1}^N \omega_i}. \end{aligned}$$

The measure  $\mathbb{P}_N$  is extended to arbitrary subsets of  $\{0, 1\}^N$  by additivity.

Finally we want to define a probability space for a countable sequence of independent tosses. It suffices to define the appropriate finitely additive measure  $P$  on the algebra  $\mathcal{C}$  of cylinder sets. This must be consistent with the probability measures just constructed on finite sequence spaces. Therefore, if  $B$  is a subset of  $\{0, 1\}^N$  and  $\pi_N$  is the projection from  $\{0, 1\}^\infty$  onto the first  $N$  tosses, we must have

$$P(\pi_N^{-1}(B)) = \mathbb{P}_N(B) = \sum_{(\omega_1, \dots, \omega_N) \in B} \prod_{i=1}^N p^{\omega_i} q^{1-\omega_i}.$$

It can be checked (exercise) that this formula consistently defined a finitely additive probability measure on  $(\{0, 1\}^\infty, \mathcal{C})$ . It extends to a countable additive probability measure on  $(\{0, 1\}^\infty, \sigma(\mathcal{C}))$  by Theorem 8. Of course this construction yields the fair coin toss space if  $p = 1/2$ .

## 1.4 Product Spaces and Independence

Independence is the concept that distinguishes probability from other branches of analysis. It just does not arise naturally, nor is it used as much, in applications of analysis to other areas. However, there is one concept from general measure theory that is closely related, namely a product of measure spaces. This section shows how to construct product spaces and how they are related to independence. It is assumed that the reader has seen the definition of a finite product of measure spaces and understands Fubini's theorem. We will review and extend this theory, but we omit proofs.

To motivate product spaces from the viewpoint of probability, suppose that we have  $N$  separate probability spaces  $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ , each modeling the outcome of a different random experiment. We would like to create one large probability space that models all  $N$  trials at once, in such a way that their outcomes are mutually independent. The outcome of running each of the trials once may be represented as a sequence  $\omega = (\omega_1, \dots, \omega_N)$ , in the the product space

$$\bigotimes_{i=1}^N \Omega_i \triangleq \Omega_1 \times \cdots \times \Omega_N,$$

Let  $\rho_i$  be the projection of  $\Omega$  onto to its  $i^{\text{th}}$  coordinate:  $\rho_i(\omega) = \omega_i$ . Then

$$\rho_i^{-1}(\mathcal{F}_i) = \{\rho_i^{-1}(B); B \in \mathcal{F}_i\}$$

is the  $\sigma$ -algebra of subsets in  $\Omega$  corresponding to events concerning only the outcome of the first trial. Therefore,

$$\bigotimes_{i=1}^N \mathcal{F}_i \triangleq \sigma\left(\bigcup_{i=1}^N \rho_i^{-1}(\mathcal{F}_i)\right)$$

is the smallest  $\sigma$ -algebra of events that contains the events pertinent to each trial; it is called the product  $\sigma$ -algebra. In mathematical language, our aim is to define a probability measure  $\mathbb{P}$  on this  $\sigma$ -algebra with the properties,

$$\mathbb{P}\left(\rho_i^{-1}(B)\right) = \mathbb{P}_i(B), \quad \text{for any } 1 \leq i \leq N \text{ and } B \in \mathcal{F}_i \quad (1.15)$$

$$\rho_1^{-1}(\mathcal{F}_1), \dots, \rho_N^{-1}(\mathcal{F}_N) \text{ are independent with respect to } \mathbb{P} \quad (1.16)$$

As usual the approach is to first define a finitely additive probability space consistent with (1.15) and (1.16). The simplest events which can involve the

outcome of more than one trial are products  $B_1 \times \cdots \times B_N$ , where  $B_i \in \mathcal{F}_i$  for each  $i$ . These products are called *measurable rectangles*. Let  $\mathcal{R}$  denote the family of all finite disjoint unions of measurable rectangles. A simple application of Theorem 5 shows that it is an algebra. Obviously, it generates the product  $\sigma$ -algebra,  $\bigotimes_{i=1}^N \mathcal{F}_i$ . The finitely additive product of  $\mathbb{P}_1, \dots, \mathbb{P}_N$  is the measure on  $\mathcal{R}$  defined by

$$\left[ \times_{i=1}^N \mathbb{P}_i \right] \left( B_1 \times \cdots \times B_N \right) \triangleq \prod_{i=1}^N \mathbb{P}_i(B_i). \quad (1.17)$$

It is extended to  $\mathcal{R}$  by finite additivity and it can be shown that this definition is consistent, in the sense that the same value is obtained no matter how an event in  $\mathcal{R}$  is expressed as a finite disjoint union of measurable rectangles. It is the only measure on  $\mathcal{R}$  consistent with the requirement (1.16) of independence among the  $\sigma$ -algebras  $\rho_i^{-1}(\mathcal{F}_i)$ , because  $B_1 \times \cdots \times B_N = \rho_1^{-1}(B_1) \cap \cdots \cap \rho_N^{-1}(B_N)$ .

**Theorem 10** *The finitely additive probability measure  $\times_{i=1}^N \mathbb{P}_i$  admits a unique extension to a probability measure on the product  $\sigma$ -algebra  $\bigotimes_{i=1}^N \mathcal{F}_i$ . This extension is denoted  $\bigotimes_{i=1}^N \mathbb{P}_i$  and is called the product probability measure. The  $\sigma$ -algebras  $\rho_1^{-1}(\mathcal{F}_1), \dots, \rho_N^{-1}(\mathcal{F}_N)$  are independent with respect to the product probability measure.*

**Proof of the independence claim.** The independence of  $\rho_1^{-1}(\mathcal{F}_1), \dots, \rho_N^{-1}(\mathcal{F}_N)$  is a direct consequence of the construction. Since the product measure extends  $\times_{i=1}^N \mathbb{P}_i$ , it follows from 1.17 that

$$\left[ \bigotimes_{i=1}^N \mathbb{P}_i \right] \left( \rho_k^{-1}(B) \right) = \left[ \bigotimes_{i=1}^N \mathbb{P}_i \right] \left( \Omega_1 \times \cdots \times \Omega_{k-1} \times B \times \Omega_{k+1} \times \cdots \times \Omega_N \right) = \mathbb{P}_k(B).$$

Then, using this and another application of (1.17),

$$\begin{aligned} \left[ \times_{i=1}^N \mathbb{P}_i \right] \left( \rho_1^{-1}(B_1) \times \cdots \times \rho_N^{-1}(B_N) \right) &= \left[ \times_{i=1}^N \mathbb{P}_i \right] \left( B_1 \times \cdots \times B_N \right) \\ &= \prod_{i=1}^N \mathbb{P}_i(B_i) = \prod_{i=1}^N \left[ \bigotimes_{i=1}^N \mathbb{P}_i \right] \left( \rho_i^{-1}(B_i) \right). \end{aligned}$$

Since  $\Omega \in \rho_i^{-1}(\mathcal{F}_i)$  for each  $i$ , an application of Lemma 6 completes the proof of independence.  $\diamond$

A proof of the validity of the extension is given at the end of this section.

It is very important to probability theory to extend the product measure construction to countable products of probability spaces. The reason is the same we gave to motivate the countable coin toss space. Countable products provide the framework for analyzing what happens in the limit as the number of repetitions of a trial increases without bound. In fact the countable coin toss space is a simple example of a countable product of identical probability spaces. We state next the general construction.

Let  $\{(\Omega_i, \mathcal{F}_i, \mathbb{P}_i); i \geq 1\}$  be a countable family of probability spaces. We follow the steps used in the construction of the infinite fair coin toss space, using the finite product measures just constructed. Define the countable product space,

$$\bigotimes_{i=1}^{\infty} \Omega_i$$

For each  $N \geq 1$ , let  $\pi_N$  be the projection of  $\bigotimes_{i=1}^{\infty} \Omega_i$  on its first  $N$  coordinates, and let

$$\mathcal{C}_N \triangleq \pi_N^{-1} \left( \bigotimes_{i=1}^N \mathcal{F}_i \right).$$

Then  $\{\mathcal{C}_N\}$  is an increasing sequence of  $\sigma$ -algebras, and their union  $\mathcal{C} \triangleq \bigcup_{N=1}^{\infty} \mathcal{C}_N$  is an algebra, called the algebra of cylinder sets. For any  $N$  and  $B \in \bigotimes_{i=1}^N \mathcal{F}_i$  define

$$P\left(\pi_N^{-1}(B)\right) = \left[ \bigotimes_{i=1}^N \mathbb{P}_i \right](B).$$

It can be checked that this is a consistent definition of a finitely additive probability measure on  $\mathcal{C}$ . Finally, define the product  $\sigma$ -algebra by

$$\bigotimes_{i=1}^{\infty} \mathcal{F}_i = \sigma(\mathcal{C}).$$

**Theorem 11**  *$P$  extends to a countably additive measure on  $\left(\bigotimes_{i=1}^{\infty} \Omega_i, \bigotimes_{i=1}^{\infty} \mathcal{F}_i\right)$  called the product measure and denoted by  $\bigotimes_{i=1}^{\infty} \mathbb{P}_i$ . For each  $k$  and  $B \in \mathcal{F}_k$ ,  $\bigotimes_{i=1}^{\infty} \mathbb{P}_i\left(\rho_k^{-1}(B)\right) = \mathbb{P}_k(B)$ , and the  $\sigma$ -algebras  $\{\rho_i^{-1}(\mathcal{F}_i); i \geq 1\}$  are independent from one another with respect to the product measure.*

The reader should check that the infinite, independent coin toss space constructed in section 1.3.2 is the product of a countable number of copies of  $(\{0, 1\}, \{\emptyset, \{0, 1\}, \{0\}, \{1\}\}, \mathbb{P}_{1,p})$ , where  $\mathbb{P}(\{x\}) = p^x(1-p)^{1-x}$  for  $x \in \{0, 1\}$ .

**Proof of extension claim of Theorem 10, sketch:** (Optional; the proof is not necessary for following the rest of the text. It will use concepts from measure-theoretic integration theory.)

We will prove the case  $N = 2$  in detail, as this contains all the ideas, and then comment on how to pass to the general case. For notational convenience, abbreviate the finitely additive product measure  $\times_{i=1}^2 \mathbb{P}_i$  by  $P$ . To show  $P$  on  $\mathcal{R}$  has a finitely additive extension we will show that it is continuous from above at  $\emptyset$ , by proving the contrapositive: if  $\{A_n\}$  is a decreasing sequence of events in  $\mathcal{R}$  and  $\lim_{n \rightarrow \infty} P(A_n) > 0$ , then  $\bigcap A_n \neq \emptyset$ .

If  $A \subset \Omega_1 \times \Omega_2$ ,

$$A(\omega_1) = \{\omega_2 \in \Omega_2; (\omega_1, \omega_2) \in A\}$$

denotes its section at  $\omega_1 \in \Omega_1$ . Let  $A \in \mathcal{R}$ , and write  $A$  as the disjoint union of measurable rectangles  $A = \bigcup_1^k U_i \times V_i$ . The section

$$A(\omega_1) = \bigcup_{i, \omega_1 \in U_i} V_i.$$

and the sets  $V_i$  in this union are disjoint. It follows that

$$\mathbb{P}_2(A(\omega_1)) = \sum_1^k \mathbf{1}_{U_i}(\omega_1) \mathbb{P}_2(V_i)$$

where  $\mathbf{1}_U$  denote the indicator function of a set  $V$ ;  $\mathbf{1}_V(x) = 1$  if  $x \in V$  and  $\mathbf{1}_V(x) = 0$  otherwise. This defines a function on  $\Omega_1$  that is measurable with respect to the  $\sigma$ -algebra  $\mathcal{F}_1$ , since it is just a linear combination of indicator functions of measurable sets. Moreover,

$$P(A) = \sum_1^k \mathbb{P}(U_i) \mathbb{P}(V_i) = \int \mathbb{P}_2(A(\omega_1)) d\mathbb{P}_1(\omega_1),$$

as a short calculation will show. Now suppose that  $\{A_n\}$  is a decreasing sequence of sets in  $\mathcal{R}$  and  $\lim_{n \rightarrow \infty} P(A_n) > 0$ . The sequence of positive functions  $\{\mathbb{P}_2(A_n(\omega_1))\}$  is also decreasing, and hence admits a pointwise limit as  $n \rightarrow \infty$ . By the dominated convergence theorem,

$$0 < \lim_{n \rightarrow \infty} P(A_n) = \int \left[ \lim_{n \rightarrow \infty} \mathbb{P}_2(A_n(\omega_1)) \right] d\mathbb{P}_1(\omega_1).$$

It follows that there exists  $\omega_1^*$  such that  $\lim_{n \rightarrow \infty} \mathbb{P}_2(A_n(\omega_1^*)) > 0$ . But then, since  $\mathbb{P}_2$  is a countably additive probability measure and therefore continuous

from above at  $\emptyset$ ,  $\bigcap_{n=1}^{\infty} A_n(\omega_2) \neq \emptyset$ . If  $\omega_2^*$  is a point in this intersection, then  $(\omega_1^*, \omega_2^*) \in A_n$  for all  $n$  and therefore  $\bigcap_{n=1}^{\infty} A_n$  is also non-empty. This completes the proof that  $P$  is continuous from above at  $\emptyset$ .

The proof can be extended to larger  $N$  by induction. Suppose we have established existence of the product probability measure for  $N-1$  spaces. Again let  $P$  denote the finitely additive product measure. Consider a set  $A$  in  $\mathcal{R}$ , written as a disjoint union of rectangles in the form,

$$A = \bigcap_1^k U_i \times V_i,$$

where, for each  $i$ ,  $V_i$  is a measurable rectangle in the space  $\bigotimes_{i=2}^N \Omega_i$ . We let  $A(\omega_1) \subset \bigotimes_{i=2}^N \Omega_i$  be the section of  $A$  at  $\omega_1 \in \Omega_1$ :

$$A(\omega_1) = \bigcup_{i, \omega_i \in U_i} V_i.$$

This is a disjoint union of measurable rectangles in  $\bigotimes_{i=2}^N \Omega_i$ , and a calculation similar that undertaken above shows,

$$P(A) = \int \left[ \bigotimes_{i=2}^N \Omega_i \right] (A(\omega_1)) d\mathbb{P}_1(\omega_1).$$

The proof now follows as before. If  $\{A_n\}$  is a decreasing sequence of sets in  $\mathcal{R}$  with  $\lim_{n \rightarrow \infty} P(A_n) > 0$ , then it follows from the monotone convergence theorem that there is a  $\omega_1^*$  such that  $\lim_{n \rightarrow \infty} [\bigotimes_{i=2}^N \Omega_i] A_n(\omega_1^*)$  is non-empty for all  $n$ . But  $\{A_n(\omega_1^*)\}$  is decreasing and so by continuity from above,  $\bigcup_{n=1}^{\infty} A_n(\omega_1^*)$  is non-empty, and hence so also is  $\bigcup_{n=1}^{\infty} A_n$ .  $\diamond$

**Proof of Theorem 11 (sketch):** Let  $\mathcal{E}$  be the collection of all subsets of the infinite product space of the form  $\pi_N^{-1}(B)$  where  $B$  is a finite disjoint union of measurable rectangles in  $\{0, 1\}^N$ . It may be checked that  $\mathcal{E}$  is an algebra,  $\mathcal{E} \subset \mathcal{C}$ , and  $\sigma(\mathcal{E}) = \sigma(\mathcal{C})$ . Hence it suffice to show that the finitely additive product measure  $P$  restricted to  $\mathcal{E}$  has an extension.

Let  $A_n$  be a decreasing sequence of sets of  $\mathcal{E}$  such that  $\lim_{n \rightarrow \infty} P(A_n) > 0$ . Without loss of generality, assume that for each  $n$ ,  $A_n = \pi_n^{-1}(B_n)$ , where  $B_n$  is a finite disjoint union of measurable rectangles. Let  $A_n(\omega_1)$  denote the section at  $\omega_1$ , Then one can show that  $\omega_1 \rightarrow [\bigotimes_2^n \mathbb{P}_i](A_n(\omega_1))$  is  $\mathcal{F}_1$  measurable, that

$$P(A_n) = \int \left[ \bigotimes_2^n \mathbb{P}_i \right] (A_n(\omega_1)) d\mathbb{P}_1(\omega_1).$$

and that  $[\otimes_2^n \mathbb{P}_i](A_n(\omega_1))$  is pointwise decreasing as  $n$  increases, for every  $\omega_1$ . It follows as before that there is an  $\omega_1^*$  such that  $A_n(\omega_1^*)$  is non-empty for all  $n$ . By applying the same argument to the decreasing sequence  $A_n(\omega_1^*)$ , with respect to the finitely additive product measure defined on  $\otimes_1^\infty \Omega_i$ , one then shows there is an  $\omega_2^*$  such that  $A_n(\omega_1^*, \omega_2^*)$  is nonempty for all  $n$ . One can continue in this manner and prove by induction that there is an infinite sequence  $(\omega_1^*, \omega_2^*, \dots)$  such that  $A_n(\omega_1^*, \dots, \omega_k^*)$  is nonempty for all  $n$  and  $k$ . Finally one uses the fact that the  $A_n$  are cylinder sets, exactly as in the proof of Theorem 7, to show that  $(\omega_1^*, \omega_2^*, \dots) \in \bigcap_1^\infty A_n$ , and hence that the intersection is nonempty.

The independence of the  $\rho_1^{-1}(\mathcal{F}_1), \rho_2^{-1}(\mathcal{F}_2), \dots$ , follows from the definition of the product measure.  $\diamond$