Chapter 1

Motivation

SECTION 1 offers some reasons for why anyone who uses probability should know about the measure theoretic approach.

SECTION 2 describes some of the added complications, and some of the compensating benefits that come with the rigorous treatment of probabilities as measures.

SECTION 3 argues that there are advantages in approaching the study of probability theory via expectations, interpreted as linear functionals, as the basic concept.

SECTION 4 describes the de Finetti convention of identifying a set with its indicator function, and of using the same symbol for a probability measure and its corresponding expectation.

*SECTION *5 presents a fair-price interpretation of probability, which emphasizes the linearity properties of expectations. The interpretation is sometimes a useful guide to intuition.*

1. Why bother with measure theory?

Following the appearance of the little book by Kolmogorov (1933), which set forth a measure theoretic foundation for probability theory, it has been widely accepted that probabilities should be studied as special sorts of measures. (More or less true—see the Notes to the Chapter.) Anyone who wants to understand modern probability theory will have to learn something about measures and integrals, but it takes surprisingly little to get started.

For a rigorous treatment of probability, the measure theoretic approach is a vast improvement over the arguments usually presented in undergraduate courses. Let me remind you of some difficulties with the typical introduction to probability.

Independence

There are various elementary definitions of independence for random variables. For example, one can require factorization of distribution functions,

$$\mathbb{P}\{X \le x, Y \le y\} = \mathbb{P}\{X \le x\}\,\mathbb{P}\{Y \le y\} \qquad \text{for all real } x, y.$$

The problem with this definition is that one needs to be able to calculate distribution functions, which can make it impossible to establish rigorously some desirable

properties of independence. For example, suppose $X_1, \ldots, X_4$ are independent random variables. How would you show that

$$Y = X_1 X_2 \left[\log\left(\frac{X_1^2 + X_2^2}{|X_1| + |X_2|} \right) + \frac{|X_1|^3 + X_2^3}{X_1^4 + X_2^4} \right]$$

is independent of

$$Z = \sin\left[X_3 + X_3^2 + X_3 X_4 + X_4^2 + \sqrt{X_3^4 + X_4^4} \right],$$

by means of distribution functions? Somehow you would need to express events $\{Y \le y, Z \le z\}$ in terms of the events $\{X_i \le x_i\}$, which is not an easy task. (If you did figure out how to do it, I could easily make up more taxing examples.)

You might also try to define independence via factorization of joint density functions, but I could invent further examples to make your life miserable, such as problems where the joint distribution of the random variables are not even given by densities. And if you could grind out the joint densities, probably by means of horrible calculations with Jacobians, you might end up with the mistaken impression that independence had something to do with the smoothness of the transformations.

The difficulty disappears in a measure theoretic treatment, as you will see in Chapter 4. Facts about independence correspond to facts about product measures.

Discrete versus continuous

Most introductory texts offer proofs of the Tchebychev inequality,

$$\mathbb{P}\{|X - \mu| \ge \epsilon\} \le \text{var}(X)/\epsilon^2,$$

where μ denotes the expected value of X. Many texts even offer two proofs, one for the discrete case and another for the continuous case. Indeed, introductory courses tend to split into at least two segments. First one establishes all manner of results for discrete random variables and then one reproves almost the same results for random variables with densities.

Unnecessary distinctions between discrete and continuous distributions disappear in a measure theoretic treatment, as you will see in Chapter 3.

Univariate versus multivariate

The unnecessary repetition does not stop with the discrete/continuous dichotomy. After one masters formulae for functions of a single random variable, the whole process starts over for several random variables. The univariate definitions acquire a prefix *joint*, leading to a whole host of new exercises in multivariate calculus: joint densities, Jacobians, multiple integrals, joint moment generating functions, and so on.

Again the distinctions largely disappear in a measure theoretic treatment. Distributions are just image measures; joint distributions are just image measures for maps into product spaces; the same definitions and theorems apply in both cases. One saves a huge amount of unnecessary repetition by recognizing the role of image

measures (described in Chapter 2) and recognizing joint distributions as measures on product spaces (described in Chapter 4).

Approximation of distributions

Roughly speaking, the central limit theorem asserts:

> *If $\xi_1, \ldots, \xi_n$ are independent random variables with zero expected values and variances summing to one, and if none of the ξ_i makes too large a contribution to their sum, then $\xi_1 + \ldots + \xi_n$ is approximately $N(0,1)$ distributed.*

What exactly does that mean? How can something with a discrete distribution, such as a standardized Binomial, be approximated by a smooth normal distribution? The traditional answer (which is sometimes presented explicitly in introductory texts) involves pointwise convergence of distribution functions of random variables; but the central limit theorem is seldom established (even in introductory texts) by checking convergence of distribution functions. Instead, when proofs are given, they typically involve checking of pointwise convergence for some sort of generating function. The proof of the equivalence between convergence in distribution and pointwise convergence of generating functions is usually omitted. The treatment of convergence in distribution for random vectors is even murkier.

As you will see in Chapter 7, it is far cleaner to start from a definition involving convergence of expectations of "smooth functions" of the random variables, an approach that covers convergence in distribution for random variables, random vectors, and even random elements of metric spaces, all within a single framework.

***

In the long run the measure theoretic approach will save you much work and help you avoid wasted effort with unnecessary distinctions.

2. The cost and benefit of rigor

In traditional terminology, probabilities are numbers in the range $[0,1]$ attached to events, that is, to subsets of a sample space Ω. They satisfy the rules

(i) $\mathbb{P}\emptyset = 0$ *and* $\mathbb{P}\Omega = 1$

(ii) for disjoint events $A_1, A_2, \ldots$, the probability of their union, $\mathbb{P}(\cup_i A_i)$, is equal to $\sum_i \mathbb{P}A_i$, the sum of the probabilities of the individual events.

When teaching introductory courses, I find that it pays to be a little vague about the meaning of the dots in (ii), explaining only that it lets us calculate the probability of an event by breaking it into disjoint pieces whose probabilities are summed. Probabilities add up in the same way as lengths, areas, volumes, and masses. The fact that we sometimes need a countable infinity of pieces (as in calculations involving potentially infinite sequences of coin tosses, for example) is best passed off as an obvious extension of the method for an arbitrarily large, finite number of pieces.

In fact the extension is not at all obvious, mathematically speaking. As explained by Hawkins (1979), the possibility of having the additivity property (ii)

hold for countable collections of disjoint events, a property known officially as ***countable additivity***, is one of the great discoveries of modern mathematics. In his 1902 doctoral dissertation, Henri Lebesgue invented a method for defining lengths of complicated subsets of the real line, in a countably additive way. The definition has the subtle feature that not every subset has a length. Indeed, under the usual axioms of set theory, it is impossible to extend the concept of length to *all* subsets of the real line while preserving countable additivity.

The same subtlety carries over to probability theory. In general, the collection of events to which countably additive probabilities are assigned cannot include all subsets of the sample space. The domain of the set function $\mathbb{P}$ (the ***probability measure***) is usually just a ***sigma-field***, a collection of subsets of Ω with properties that will be defined in Chapter 2.

Many probabilistic ideas are greatly simplified by reformulation as properties of sigma-fields. For example, the unhelpful multitude of possible definitions for independence coalesce nicely into a single concept of independence for sigma-fields.

The sigma-field limitation turns out to be less of a disadvantage than might be feared. In fact, it has positive advantages when we wish to prove some probabilistic fact about all events in some sigma-field, $\mathcal{A}$. The obvious line of attack—first find an explicit representation for the typical member of $\mathcal{A}$, then check the desired property directly—usually fails. Instead, as you will see in Chapter 2, an indirect approach often succeeds.

(a) Show directly that the desired property holds for all events in some subclass $\mathcal{E}$ of "simpler sets" from $\mathcal{A}$.

(b) Show that $\mathcal{A}$ is the smallest sigma-field for which $\mathcal{A} \supseteq \mathcal{E}$.

(c) Show that the desired property is preserved under various set theoretic operations. For example, it might be possible to show that if two events have the property then so does their union.

(d) Deduce from (c) that the collection $\mathcal{B}$ of all events with the property forms a sigma-field of subsets of Ω. That is, $\mathcal{B}$ is a sigma-field, which, by (a), has the property $\mathcal{B} \supseteq \mathcal{E}$.

(e) Conclude from (b) and (d) that $\mathcal{B} \supseteq \mathcal{A}$. That is, the property holds for all members of $\mathcal{A}$.

> REMARK. Don't worry about the details for the moment. I include the outline in this Chapter just to give the flavor of a typical measure theoretic proof. I have found that some students have trouble adapting to this style of argument.

The indirect argument might seem complicated, but, with the help of a few key theorems, it actually becomes routine. In the literature, it is not unusual to see applications abbreviated to a remark like "a simple generating class argument shows . . . ," with the reader left to fill in the routine details.

Lebesgue applied his definition of length (now known as Lebesgue measure) to the construction of an integral, extending and improving on the Riemann integral. Subsequent generalizations of Lebesgue's concept of measure (as in the 1913 paper of Radon and other developments described in the Epilogue to

Hawkins 1979) eventually opened the way for Kolmogorov to identify probabilities with measures on sigma-fields of events on general sample spaces. From the Preface to Kolmogorov (1933), in the 1950 translation by Morrison:

> The purpose of this monograph is to give an axiomatic foundation for the theory of probability. The author set himself the task of putting in their natural place, among the general notions of modern mathematics, the basic concepts of probability theory—concepts which until recently were considered to be quite peculiar.
>
> This task would have been a rather hopeless one before the introduction of Lebesgue's theories of measure and integration. However, after Lebesgue's publication of his investigations, the analogies between measure of a set and probability of an event, and between integral of a function and mathematical expectation of a random variable, became apparent. These analogies allowed of further extensions; thus, for example, various properties of independent random variables were seen to be in complete analogy with the corresponding properties of orthogonal functions. But if probability theory was to be based on the above analogies, it still was necessary to make the theories of measure and integration independent of the geometric elements which were in the foreground with Lebesgue. This has been done by Fréchet.
>
> While a conception of probability theory based on the above general viewpoints has been current for some time among certain mathematicians, there was lacking a complete exposition of the whole system, free of extraneous complications. (Cf., however, the book by Fréchet ...)

Kolmogorov identified random variables with a class of real-valued functions (the ***measurable functions***) possessing properties allowing them to coexist comfortably with the sigma-field. Thereby he was also able to identify the expectation operation as a special case of integration with respect to a measure. For the newly restricted class of random variables, in addition to the traditional properties

(i) $\mathbb{E}(c_1X_1 + c_2X_2) = c_1\mathbb{E}(X_1) + c_2\mathbb{E}(X_2)$, for constants c_1 and c_2,

(ii) $\mathbb{E}(X) \geq \mathbb{E}(Y)$ if $X \geq Y$,

he could benefit from further properties implied by the countable additivity of the probability measure.

As with the sigma-field requirement for events, the measurability restriction on the random variables came with benefits. In modern terminology, no longer was $\mathbb{E}$ just an ***increasing linear functional*** on the space of real random variables (with some restrictions to avoid problems with infinities), but also it had acquired some continuity properties, making possible a rigorous treatment of limiting operations in probability theory.

3. Where to start: probabilities or expectations?

From the example set by Lebesgue and Kolmogorov, it would seem natural to start with probabilities of events, then extend, via the operation of integration, to the study of expectations of random variables. Indeed, in many parts of the mathematical world that is the way it goes: probabilities are the basic quantities, from which expectations of random variables are derived by various approximation arguments.

The apparently natural approach is by no means the only possibility, as anyone brought up on the works of the fictitious French author Bourbaki could affirm. (The treatment of measure theory, culminating with Bourbaki 1969, started from integrals defined as linear functionals on appropriate spaces of functions.) Moreover, historically speaking, expectation has a strong claim to being the preferred starting point for a theory of probability. For instance, in his discussion of the 1657 book *Calculating in Games of Chance* by Christian Huygens, Hacking (1978, page 97) commented:

> The fair prices worked out by Huygens are just what we would call the expectations of the corresponding gambles. His approach made expectation a more basic concept than probability, and this remained so for about a century.

The fair price interpretation is sketched in Section 5.

The measure theoretic history of integrals as linear functionals also extends back to the early years of the twentieth century, starting with Daniell (1918), who developed a general theory of integration via extension of linear functionals from small spaces of functions to larger spaces. It is also significant that, in one of the greatest triumphs of measure theory, Wiener (1923, Section 10) defined what is now known as Wiener measure (thereby providing a rigorous basis for the mathematical theory of Brownian motion) as an averaging operation for functionals defined on Brownian motion paths, citing Daniell (1919) for the basic extension theorem.

There are even better reasons than historical precedent for working with expectations as the basic concept. Whittle (1992), in the *Preface* to an elegant, intermediate level treatment of *Probability via Expectations*, presented some arguments:

(i) To begin with, people probably have a better intuition for what is meant by an 'average value' than for what is meant by a 'probability.'

(ii) Certain important topics, such as optimization and approximation problems, can be introduced and treated very quickly, just because they are phrased in terms of expectations.

(iii) Most elementary treatments are bedeviled by the apparent need to ring the changes of a particular proof or discussion for all the special cases of continuous or discrete distribution, scalar or vector variables, etc. In the expectations approach these are indeed seen as special cases, which can be treated with uniformity and economy.

His list continued. I would add that:

(a) It is often easier to work with the linearity properties of integrals than with the additivity properties of measures. For example, many useful probability inequalities are but thinly disguised consequences of pointwise inequalities, translated into probability form by the linearity and increasing properties of expectations.

(b) The linear functional approach, via expectations, can save needless repetition of arguments. Some theorems about probability measures, as set functions, are just special cases of more general results about expectations.

(c) When constructing new probability measures, we save work by defining the integral of measurable functions directly, rather than passing through the preliminary step of building the set function then establishing theorems about the corresponding integrals. As you will see repeatedly, definitions and theorems sometimes collapse into a single operation when expressed directly in terms of expectations, or integrals.

***

I will explain the essentials of measure theory in Chapter 2, starting from the traditional set-function approach but working as quickly as I can towards systematic use of expectations.

4. The de Finetti notation

The advantages of treating expectation as the basic concept are accentuated by the use of an elegant notation strongly advocated by de Finetti (1972, 1974). Knowing that many traditionally trained probabilists and statisticians find the notation shocking, I will introduce it slowly, in an effort to explain why it is worth at least a consideration. (Immediate enthusiastic acceptance is more than I could hope for.)

Ordinary algebra is easier than Boolean algebra. The correspondence $A \leftrightarrow \mathbb{I}_A$ between subsets A of a fixed set $\mathcal{X}$ and their indicator functions,

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

transforms Boolean algebra into ordinary pointwise algebra with functions. I claim that probability theory becomes easier if one works systematically with expectations of indicator functions, $\mathbb{E}\mathbb{I}_A$, rather than with the corresponding probabilities of events.

Let me start with the assertions about algebra and Boolean algebra. The operations of union and intersection correspond to pointwise maxima (denoted by max or the symbol $\vee$) and pointwise minima (denoted by min or the symbol $\wedge$), or pointwise products:

$$\mathbb{I}_{\cup_i A_i}(x) = \bigvee_i \mathbb{I}_{A_i}(x) \qquad \text{and} \qquad \mathbb{I}_{\cap_i A_i}(x) = \bigwedge_i \mathbb{I}_{A_i}(x) = \prod_i \mathbb{I}_{A_i}(x).$$

Complements correspond to subtraction from one: $\mathbb{I}_{A^c}(x) = 1 - \mathbb{I}_A(x)$. Derived operations, such as the set theoretic difference $A \backslash B := A \cap B^c$ and the symmetric difference, $A \Delta B := (A \backslash B) \cup (B \backslash A)$, also have simple algebraic counterparts:

$$\begin{aligned} \mathbb{I}_{A\backslash B}(x) &= (\mathbb{I}_A(x) - \mathbb{I}_B(x))^+ := \max\left(0, \mathbb{I}_A(x) - \mathbb{I}_B(x)\right), \\ \mathbb{I}_{A\Delta B}(x) &= |\mathbb{I}_A(x) - \mathbb{I}_B(x)|. \end{aligned}$$

To check these identities, just note that the functions take only the values 0 and 1, then determine which combinations of indicator values give a 1. For example, $|\mathbb{I}_A(x) - \mathbb{I}_B(x)|$ takes the value 1 when exactly one of $\mathbb{I}_A(x)$ and $\mathbb{I}_B(x)$ equals 1.

The algebra looks a little cleaner if we omit the argument x. For example, the horrendous set theoretic relationship

$$\left(\cap_{i=1}^n A_i\right) \Delta \left(\cap_{i=1}^n B_i\right) \subseteq \cup_{i=1}^n \left(A_i \Delta B_i\right)$$

corresponds to the pointwise inequality

$$\left|\prod_i \mathbb{I}_{A_i} - \prod_i \mathbb{I}_{B_i}\right| \le \max_i \left|\mathbb{I}_{A_i} - \mathbb{I}_{B_i}\right|,$$

whose verification is easy: when the right-hand side takes the value 1 the inequality is trivial, because the left-hand side can take only the values 0 or 1; and when right-hand side takes the value 0, we have $\mathbb{I}_{A_i} = \mathbb{I}_{B_i}$ for all i, which makes the left-hand side zero.

<1> **Example.** One could establish an identity such as

$$(A \Delta B) \Delta (C \Delta D) = A \Delta \left(B \Delta (C \Delta D)\right)$$

by expanding both sides into a union of many terms. It is easier to note the pattern for indicator functions. The set $A \Delta B$ is the region where $\mathbb{I}_A + \mathbb{I}_B$ takes an odd value (that is, the value 1); and $(A \Delta B) \Delta C$ is the region where $(\mathbb{I}_A + \mathbb{I}_B) + \mathbb{I}_C$ takes an odd value. And so on. In fact both sides of the set theoretic identity equal the region where $\mathbb{I}_A + \mathbb{I}_B + \mathbb{I}_C + \mathbb{I}_D$ takes an odd value. Associativity of set theoretic differences is a consequence of associativity of pointwise addition. □

<2> **Example.** The lim sup of a sequence of sets $\{A_n : n \in \mathbb{N}\}$ is defined as

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{i \ge n} A_i .$$

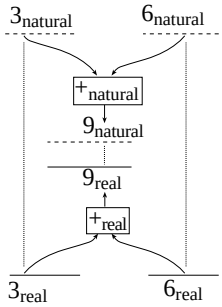
That is, the lim sup consists of those x for which, to each n there exists an $i \ge n$ such that $x \in A_i$. Equivalently, it consists of those x for which $x \in A_i$ for infinitely many i. In other words,

$$\mathbb{I}_{\limsup_n A_n} = \limsup_n \mathbb{I}_{A_n} .$$

Do you really need to learn the new concept of the lim sup of a sequence of sets? Theorems that work for lim sups of sequences of functions automatically carry over to theorems about sets. There is no need to prove everything twice. The correspondence between sets and their indicators saves us from unnecessary work. □

After some repetition, it becomes tiresome to have to keep writing the $\mathbb{I}$ for the indicator function. It would be much easier to write something like $\tilde{A}$ in place of $\mathbb{I}_A$. The indicator of the lim sup of a sequence of sets would then be written $\limsup_n \tilde{A}_n$, with only the tilde to remind us that we are referring to functions. But why do we need reminding? As the example showed, the concept for the lim sup of sets is really just a special case of the concept for sequences of functions. Why preserve a distinction that hardly matters?

There is a well established tradition in Mathematics for choosing notation that eliminates inessential distinctions. For example, we use the same symbol 3 for the natural number and the real number, writing $3 + 6 = 9$ as an assertion both about addition of natural numbers and about addition of real numbers.



It does not matter if we cannot tell immediately which interpretation is intended, because we know there is a one-to-one correspondence between natural numbers and a subset of the real numbers, which preserves all the properties of interest. Formally, there is a map $\psi : \mathbb{N} \to \mathbb{R}$ for which

$$\psi(x +_{natural} y) = \psi(x) +_{real} \psi(y) \qquad \text{for all } x, y \text{ in } \mathbb{N},$$

with analogous equalities for other operations. (Notice that I even took care to distinguish between addition as a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ and as a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$.) The map ψ is an isomorphism between $\mathbb{N}$ and a subset of $\mathbb{R}$.

> REMARK. Of course there are some situations where we need to distinguish between a natural number and its real counterpart. For example, it would be highly confusing to use indistinguishable symbols when first developing the properties of the real number system from the properties of the natural numbers. Also, some computer languages get very upset when a function that expects a floating point argument is fed an integer variable; some languages even insist on an explicit conversion between types.

We are faced with a similar overabundance of notation in the correspondence between sets and their indicator functions. Formally, and traditionally, we have a map $A \mapsto \mathbb{I}_A$ from sets into a subset of the nonnegative real functions. The map preserves the important operations. It is firmly in the Mathematical tradition that we should follow de Finetti's suggestion and **use the same symbol for a set and its indicator function**.

> REMARK. A very similar convention has been advocated by the renowned computer scientist, Donald Knuth, in an expository article (Knuth 1992). He attributed the idea to Kenneth Iversen, the inventor of the programming language APL.

In de Finetti's notation the assertion from Example <2> becomes

$$\limsup A_n = \limsup A_n,$$

a fact that is quite easy to remember. The theorem about lim sups of sequences of sets has become incorporated into the notation; we have one less theorem to remember.

The second piece of de Finetti notation is suggested by the same logic that encourages us to replace $+_{natural}$ and $+_{real}$ by the single addition symbol: use the same symbol when extending the domain of definition of a function. For example, the symbol "sin" denotes both the function defined on the real line and its extension to the complex domain. More generally, if we have a function g with domain G_0, which can be identified with a subset $\widetilde{G}_0$ of some $\widetilde{G}$ via a correspondence $x \leftrightarrow \tilde{x}$, and if $\tilde{g}$ is a function on $\widetilde{G}$ for which $\tilde{g}(\tilde{x}) = g(x)$ for x in G_0, then why not write g instead of $\tilde{g}$ for the function with the larger domain?

With probability theory we often use $\mathbb{P}$ to denote a probability measure, as a map from a class $\mathcal{A}$ (a sigma-field) of subsets of some Ω into the subinterval $[0, 1]$ of the real line. The correspondence $A \leftrightarrow \tilde{A} := \mathbb{I}_A$, between a set A and its indicator function $\tilde{A}$, establishes a correspondence between $\mathcal{A}$ and a subset of the collection of

random variables on Ω. The expectation maps random variables into real numbers, in such a way that $\mathbb{E}(\tilde{A}) = \mathbb{P}(A)$. This line of thinking leads us to de Finetti's second suggestion: **use the same symbol for expectation and probability measure**, writing $\mathbb{P}X$ instead of $\mathbb{E}X$, and so on.

The de Finetti notation has an immediate advantage when we deal with several probability measures, $\mathbb{P}$, $\mathbb{Q}$, ... simultaneously. Instead of having to invent new symbols $\mathbb{E}_{\mathbb{P}}$, $\mathbb{E}_{\mathbb{Q}}$, ..., we reuse $\mathbb{P}$ for the expectation corresponding to $\mathbb{P}$, and so on.

> REMARK. You might have the concern that you will not be able to tell whether $\mathbb{P}A$ refers to the probability of an event or the expected value of the corresponding indicator function. The ambiguity should not matter. Both interpretations give the same number; you will never be faced with a choice between two different values when choosing an interpretation. If this ambivalence worries you, I would suggest going systematically with the expectation/indicator function interpretation. It will never lead you astray.

<3> **Example.** For a finite collection of events $A_1, \ldots, A_n$, the so-called ***method of inclusion and exclusion*** asserts that the probability of the union $\cup_{i\le n} A_i$ equals

$$\sum_i \mathbb{P}A_i - \sum_{i\neq j} \mathbb{P}(A_i \cap A_j) + \sum_{i,j,k} \{i, j, k \text{ distinct}\} \mathbb{P}(A_i \cap A_j \cap A_k) - \ldots \pm \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n).$$

The equality comes by taking expectations on both sides of an identity for (indicator) functions,

$$\cup_{i\le n} A_i = \sum_i A_i - \sum_{i\neq j} A_i A_j + \sum_{i,j,k} \{i, j, k \text{ distinct}\} A_i A_j A_k - \ldots \pm A_1 A_2 \ldots A_n.$$

The right-hand side of this identity is just the expanded version of $1 - \prod_{i\le n}(1 - A_i)$. The identity is equivalent to

$$1 - \cup_{i\le n} A_i = \prod_{i\le n} (1 - A_i)\,,$$

which presents two ways of expressing the indicator function of $\cap_{i\le n} A_i^c$. See Problem [1] for a generalization. □

<4> **Example.** Consider Tchebychev's inequality, $\mathbb{P}\{|X - \mu| \ge \epsilon\} \le \text{var}(X)/\epsilon^2$, for each $\epsilon > 0$, and each random variable X with expected value $\mu := \mathbb{P}X$ and finite variance, $\text{var}(X) := \mathbb{P}(X - \mu)^2$. On the left-hand side of the inequality we have the probability of an event. Or is it the expectation of an indicator function? Either interpretation is correct, but the second is more helpful. The inequality is a consequence of the increasing property for expectations invoked for a pair of functions, $\{|X - \mu| \ge \epsilon\} \le (X - \mu)^2/\epsilon^2$. The indicator function on the left-hand side takes only the values 0 and 1. The quadratic function on the right-hand side is nonnegative, and is ≥ 1 whenever the left-hand side equals 1. □

***

For the remainder of the book, I will be using the same symbol for a set and its indicator function, and writing $\mathbb{P}$ instead of $\mathbb{E}$ for expectation.

> REMARK. For me, the most compelling reason to adopt the de Finetti notation, and work with $\mathbb{P}$ as a linear functional defined for random variables, was not that I would save on symbols, nor any of the other good reasons listed at the end of

Section 3. Instead, I favor the notation because, once the initial shock of seeing old symbols used in new ways wore off, it made probability theory easier. I can truly claim to have gained better insight into classical techniques through the mere fact of translating them into the new notation. I even find it easier to invent new arguments when working with a notation that encourages thinking in terms of linearity, and which does not overemphasize the special role for expectations of functions that take only the values 0 and 1 by according them a different symbol.

The hope that I might convince probability users of some of the advantages of de Finetti notation was, in fact, one of my motivations for originally deciding to write yet another book about an old subject.

*5. Fair prices

For the understanding of this book the interpretation of probability as a model for uncertainty is not essential. You could study it purely as a piece of mathematics, divorced from any interpretation but then you would forgo much of the intuition that accompanies the various interpretations.

The most widely accepted view interprets probabilities and expectations as long run averages, anticipating the formal laws of large numbers that make precise a sense in which averages should settle down to expectations over a long sequence of independent trials. As an aid to intuition I also like another interpretation, which does not depend on a preliminary concept of independence, and which concentrates attention on the linearity properties of expectations.

Consider a situation—a bet if you will–where you stand to receive an uncertain return X. You could think of X as a random variable, a real-valued function on a set Ω. For the moment forget about any probability measure on Ω. Suppose you consider $p(X)$ to be the fair price to pay now in order to receive X at some later time. (By *fair* I mean that you should be prepared to take either side of the bet. In particular, you should be prepared to accept a payment $p(X)$ from me now in return for giving me an amount X later.) What properties should $p(\cdot)$ have?

REMARK. As noted in Section 3, the value $p(X)$ corresponds to an expected value of the random variable X. If you already know about the possibility of infinite expectations, you will realize that I would have to impose some restrictions on the class of random variables for which fair prices are defined, if I were seriously trying to construct a rigorous system of axioms. It would suffice to restrict the argument to bounded random variables.

Your net return will be the random quantity $X'(\omega) := X(\omega) - p(X)$. Call the random variable X' a ***fair return***, the net return from a fair trade. Unless you start worrying about utilities—in which case you might consult Savage (1954) or Ferguson (1967, Section 1.4)—you should find the following properties reasonable.

(i) ***fair* + *fair* = *fair***. That is, if you consider $p(X)$ fair for X and $p(Y)$ fair for Y then you should be prepared to make both bets, paying $p(X) + p(Y)$ to receive $X + Y$.

(ii) ***constant* $\times$ *fair* = *fair***. That is, you shouldn't object if I suggest you pay $2p(X)$ to receive $2X$ (actually, that particular example is a special case of (i))

or $3.76p(X)$ to receive $3.76X$, or $-p(X)$ to receive $-X$. The last example corresponds to willingness to take either side of a fair bet. In general, to receive cX you should pay $cp(X)$, for constant c.

Properties (i) and (ii) imply that the collection of all fair returns is a vector space.

There is a third reasonable property that goes by several names: ***coherency*** or ***nonexistence of a Dutch book***, the ***no-arbitrage requirement***, or the ***no-free-lunch principle***:

(iii) There is no fair return X' for which $X'(\omega) \geq 0$ for all ω, with strict inequality for at least one ω.

(Students of decision theory might be reminded of the the concept of admissibility.) If you were to declare such an X' to be fair I would be delighted to offer you the opportunity to receive a net return of $-10^{100}X'$. I couldn't lose.

<5> **Lemma.** *Properties (i), (ii), and (iii) imply that $p(\cdot)$ is an increasing linear functional on random variables. The fair returns are those random variables for which $p(X) = 0$.*

Proof. For constants α and β, and random variables X and Y with fair prices $p(X)$ and $p(Y)$, consider the combined effect of the following fair bets:

you pay me $\alpha p(X)$ to receive αX

you pay me $\beta p(Y)$ to receive βY

I pay you $p(\alpha X + \beta Y)$ to receive $(\alpha X + \beta Y)$.

Your net return is a constant,

$$c = p(\alpha X + \beta Y) - \alpha p(X) - \beta p(Y).$$

If $c > 0$ you violate (iii); if $c < 0$ take the other side of the bet to violate (iii). That proves linearity.

To prove that $p(\cdot)$ is increasing, suppose $X(\omega) \geq Y(\omega)$ for all ω. If you claim that $p(X) < p(Y)$ then I would be happy for you to accept the bet that delivers

$$(Y - p(Y)) - (X - p(X)) = -(X - Y) - (p(Y) - p(X)),$$

which is always < 0.

If both X and $X - p(X)$ are considered fair, then the constant return $p(X) = X - (X - p(X))$ is fair, which would contradict (iii) unless $p(X) = 0$. □

As a special case, consider the bet that returns 1 if an event F occurs, and 0 otherwise. If you identify the event F with the random variable taking the value 1 on F and 0 on F^c (that is, the indicator of the event F), then it follows directly from Lemma <5> that $p(\cdot)$ is additive: $p(F_1 \cup F_2) = p(F_1) + p(F_2)$ for disjoint events F_1 and F_2. That is, p defines a finitely additive set-function on events. The set function $p(\cdot)$ has most of the properties required of a probability measure. As an exercise you might show that $p(\emptyset) = 0$ and $p(\Omega) = 1$.

Contingent bets

Things become much more interesting if you are prepared to make a bet to receive an amount X but only when some event F occurs. That is, the bet is made ***contingent***

on the occurrence of F. Typically, knowledge of the occurrence of F should change the fair price, which we could denote by $p(X \mid F)$. Expressed more compactly, the bet that returns $(X - p(X \mid F))\,F$ is fair. The indicator function F ensures that money changes hands only when F occurs.

<6> **Lemma.** *If Ω is partitioned into disjoint events $F_1, \ldots, F_k$, and X is a random variable, then $p(X) = \sum_{i=1}^{k} p(F_i)p(X \mid F_i)$.*

Proof. For a single F_i, argue by linearity that

$$0 = p\,(XF_i - p(X \mid F_i)F_i) = p(XF_i) - p(X \mid F_i)p(F_i).$$

Sum over i, using linearity again, together with the fact that $X = \sum_i XF_i$, to deduce that $p(X) = \sum_i p(XF_i) = \sum_i p(F_i)p(X \mid F_i)$, as asserted. □

Why should we restrict the Lemma to finite partitions? If we allowed countable partitions we would get the countable additivity property—the key requirement in the theory of measures. I would be suspicious of such an extension of the simple argument for finite partitions. It makes a tacit assumption that a combination of countably many fair bets is again fair. If we accept that assumption, then why not accept that arbitrary combinations of fair events are fair? For uncountably infinite collections we would run into awkward contradictions. For example, suppose ω is generated from a uniform distribution on $[0, 1]$. Let X_t be the random variable that returns 1 if $\omega = t$ and 0 otherwise. By symmetry one might expect $p(X_t) = c$ for some constant c that doesn't depend on t. But there can be no c for which

$$1 = p(1) = p\left(\textstyle\sum_{0\le t\le 1} X_t\right) \stackrel{?}{=} \sum_{0\le t\le 1} p(X_t) = \begin{cases} 0 & \text{if } c = 0 \\ \pm\infty & \text{if } c \neq 0 \end{cases}$$

Perhaps our intuition about the infinite rests on shaky analogies with the finite.

REMARK. I do not insist that probabilities must be interpreted as fair prices, just as I do not accept that all probabilities must be interpreted as assertions about long run frequencies. It is convenient that both interpretations lead to almost the same mathematical formalism. You are free to join either camp, or both, and still play by the same probability rules.

6. Problems

[1] Let $A_1, \ldots, A_N$ be events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each subset J of $\{1, 2, \ldots, N\}$ write A_J for $\cap_{i\in J} A_i$. Define $S_k := \sum_{|J|=k} \mathbb{P}A_J$, where $|J|$ denotes the number of indices in J. For $0 \le m \le N$ show that the probability $\mathbb{P}\{\text{exactly } m \text{ of the } A_i\text{'s occur}\}$ equals $\binom{m}{m}S_m - \binom{m+1}{m}S_{m+1} + \ldots \pm \binom{N}{m}S_N$. Hint: For a dummy variable z, show that $\prod_{i=1}^{N}(A_i^c + zA_i) = \sum_{k=0}^{n}\sum_{|J|=k}(z-1)^k A_J$. Expand the left-hand side, take expectations, then interpret the coefficient of z^m.

[2] Rederive the assertion of Lemma <6> by consideration of the net return from the following system of bets: (i) for each i, pay $c_i\,p(F_i)$ in order to receive c_i if F_i occurs, where $c_i := p(X \mid F_i)$; (ii) pay $-p(X)$ in order to receive $-X$; (iii) for each i, make a bet contingent on F_i, paying c_i (if F_i occurs) to receive X.

[3] For an increasing sequence of events $\{A_n : n \in \mathbb{N}\}$ with union A, show $\mathbb{P}A_n \uparrow \mathbb{P}A$.

7. Notes

See Dubins & Savage (1964) for an illustration of what is possible in a theory of probability without countable additivity.

The ideas leading up to Lebesgue's creation of his integral are described in fascinating detail in the excellent book of Hawkins (1979), which has been the starting point for most of my forays into the history of measure theory. Lebesgue first developed his new definition of the integral for his doctoral dissertation (Lebesgue 1902), then presented parts of his theory in the 1902–1903 Peccot course of lectures (Lebesgue 1904). The 1928 revision of the 1904 volume greatly expanded the coverage, including a treatment of the more general (Lebesgue-)Stieltjes integral. See also Lebesgue (1926), for a clear description of some of the ideas involved in the development of measure theory, and the *Note Historique* of Bourbaki (1969), for a discussion of later developments.

Of course it is a vast oversimplification to imagine that probability theory abruptly became a specialized branch of measure theory in 1933. As Kolmogorov himself made clear, the crucial idea was the measure theory of Lebesgue. Kolmogorov's little book was significant not just for "putting in their natural place, among the general notions of modern mathematics, the basic concepts of probability theory", but also for adding new ideas, such as probability distributions in infinite dimensional spaces (reinventing results of Daniell 1919) and a general theory of conditional probabilities and conditional expectations.

Measure theoretic ideas were used in probability theory well before 1933. For example, in the *Note* at the end of Lévy (1925) there was a clear statement of the countable additivity requirement for probabilities, but Lévy did not adopt the complete measure theoretic formalism; and Khinchin & Kolmogorov (1925) explicitly constructed their random variables as functions on $[0, 1]$, in order to avail themselves of the properties of Lebesgue measure.

It is also not true that acceptance of the measure theoretic foundation was total and immediate. For example, eight years after Kolmogorov's book appeared, von Mises (1941, page 198) asserted (emphasis in the original):

> In recapitulating this paragraph I may say: First, the axioms of Kolmogorov are concerned with the distribution function within one kollektiv and are *supplementary to my theory, not a substitute for it.* Second, using the notion of measure zero in an absolute way without reference to the arbitrarily assumed measure system, *leads to essential inconsistencies.*

See also the argument for the measure theoretic framework in the accompanying paper by Doob (1941), and the comments by both authors that follow (von Mises & Doob 1941).

For more about Kolmogorov's pivotal role in the history of modern probability, see: Shiryaev (2000), and the other articles in the same collection; the memorial

articles in the *Annals of Probability*, volume 17 (1989); and von Plato (1994), which also contains discussions of the work of von Mises and de Finetti.

References

Bourbaki, N. (1969), *Intégration sur les espaces topologiques séparés*, Éléments de mathématique, Hermann, Paris. Fascicule XXXV, Livre VI, Chapitre IX.

Daniell, P. J. (1918), 'A general form of integral', *Annals of Mathematics (series 2)* **19**, 279–294.

Daniell, P. J. (1919), 'Functions of limited variation in an infinite number of dimensions', *Annals of Mathematics (series 2)* **21**, 30–38.

de Finetti, B. (1972), *Probability, Induction, and Statistics*, Wiley, New York.

de Finetti, B. (1974), *Theory of Probability*, Wiley, New York. First of two volumes translated from *Teoria Delle probabilità*, published 1970. The second volume appeared under the same title in 1975.

Doob, J. L. (1941), 'Probability as measure', *Annals of Mathematical Statistics* **12**, 206–214.

Dubins, L. & Savage, L. (1964), *How to Gamble if You Must*, McGraw-Hill.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, Boston.

Fréchet, M. (1915), 'Sur l'intégrale d'une fonctionnelle étendue à un ensemble abstrait', *Bull. Soc. Math. France* **43**, 248–265.

Hacking, I. (1978), *The Emergence of Probability*, Cambridge University Press.

Hawkins, T. (1979), *Lebesgue's Theory of Integration: Its Origins and Development*, second edn, Chelsea, New York.

Khinchin, A. Y. & Kolmogorov, A. (1925), 'Über Konvergenz von Reihen, deren Glieder durch den Zufall bestimmt werden', *Mat. Sbornik* **32**, 668–677.

Knuth, D. E. (1992), 'Two notes on notation', *American Mathematical Monthly* **99**, 403–422.

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.

Lebesgue, H. (1902), *Intégrale, longueur, aire*. Doctoral dissertation, submitted to Faculté des Sciences de Paris. Published separately in Ann. Mat. Pura Appl. 7. Included in the first volume of his *Œuvres Scientifiques*, published in 1972 by L'Enseignement Mathématique.

Lebesgue, H. (1904), *Leçons sur l'integration et la recherche des fonctions primitives*, first edn, Gauthier-Villars, Paris. Included in the second volume of his *Œuvres Scientifiques*, published in 1972 by L'Enseignement Mathématique. Second edition published in 1928. Third edition, 'an unabridged reprint of the second edition, with minor changes and corrections', published in 1973 by Chelsea, New York.

Lebesgue, H. (1926), 'Sur le développement de la notion d'intégrale', *Matematisk Tidsskrift B*. English version in the book *Measure and Integral*, edited and translated by Kenneth O. May.

Lévy, P. (1925), *Calcul des Probabilités*, Gauthier-Villars, Paris.

Radon, J. (1913), 'Theorie und Anwendungen der absolut additiven Mengenfunctionen', *Sitzungsberichten der Kaiserlichen Akademie der Wissenschaften in Wien. Mathematisch-naturwissenschaftliche Klasse* **122**, 1295–1438.

Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, New York. Second edition, Dover, New York, 1972.

Shiryaev, A. N. (2000), Andreĭ Nikolaevich Kolmogorov: a biographical sketch of his life and creative paths, *in* 'Kolmogorov in Perspective', American Mathematical Society/London Mathematical Society.

von Mises, R. (1941), 'On the foundations of probability and statistics', *Annals of Mathematical Statistics* **12**, 191–205.

von Mises, R. & Doob, J. L. (1941), 'Discussion of papers on probability theory', *Annals of Mathematical Statistics* **12**, 215–217.

von Plato, J. (1994), *Creating Modern Probability: its Mathematics, Physics and Philosophy in Historical Perspective*, Cambridge University Press.

Whittle, P. (1992), *Probability via Expectation*, third edn, Springer-Verlag, New York. First edition 1970, under the title "Probability".

Wiener, N. (1923), 'Differential-space', *Journal of Mathematics and Physics* **2**, 131–174. Reprinted in *Selected papers of Norbert Wiener*, MIT Press, 1964.

Chapter 2

A modicum of measure theory

SECTION 1 defines measures and sigma-fields.

SECTION 2 defines measurable functions.

SECTION 3 defines the integral with respect to a measure as a linear functional on a cone of measurable functions. The definition sidesteps the details of the construction of integrals from measures.

*SECTION *4 constructs integrals of nonnegative measurable functions with respect to a countably additive measure.*

SECTION 5 establishes the Dominated Convergence theorem, the Swiss Army knife of measure theoretic probability.

SECTION 6 collects together a number of simple facts related to sets of measure zero.

*SECTION *7 presents a few facts about spaces of functions with integrable pth powers, with emphasis on the case p=2, which defines a Hilbert space.*

SECTION 8 defines uniform integrability, a condition slightly weaker than domination. Convergence in $\mathcal{L}^1$ is characterized as convergence in probability plus uniform integrability.

SECTION 9 defines the image measure, which includes the concept of the distribution of a random variable as a special case.

SECTION 10 explains how generating class arguments, for classes of sets, make measure theory easy.

*SECTION *11 extends generating class arguments to classes of functions.*

1. Measures and sigma-fields

As promised in Chapter 1, we begin with measures as set functions, then work quickly towards the interpretation of integrals as linear functionals. Once we are past the purely set-theoretic preliminaries, I will start using the de Finetti notation (Section 1.4) in earnest, writing the same symbol for a set and its indicator function.

Our starting point is a ***measure space***: a triple $(\mathcal{X}, \mathcal{A}, \mu)$, with $\mathcal{X}$ a set, $\mathcal{A}$ a class of subsets of $\mathcal{X}$, and μ a function that attaches a nonnegative number (possibly $+\infty$) to each set in $\mathcal{A}$. The class $\mathcal{A}$ and the set function μ are required to have properties that facilitate calculations involving limits along sequences.

<1> **Definition.** *Call a class $\mathcal{A}$ a sigma-field of subsets of $\mathcal{X}$ if:*

(i) the empty set $\emptyset$ and the whole space $\mathcal{X}$ both belong to $\mathcal{A}$;

(ii) if A belongs to $\mathcal{A}$ then so does its complement A^c;

(iii) if $A_1, A_2, \ldots$ is a countable collection of sets in $\mathcal{A}$ then both the union $\cup_i A_i$ and the intersection $\cap_i A_i$ are also in $\mathcal{A}$.

Some of the requirements are redundant as stated. For example, once we have $\emptyset \in \mathcal{A}$ then (ii) implies $\mathcal{X} \in \mathcal{A}$. When we come to establish properties about sigma-fields it will be convenient to have the list of defining properties pared down to a minimum, to reduce the amount of mechanical checking. The theorems will be as sparing as possible in the amount the work they require for establishing the sigma-field properties, but for now redundancy does not hurt.

The collection $\mathcal{A}$ need not contain every subset of $\mathcal{X}$, a fact forced upon us in general if we want μ to have the properties of a countably additive measure.

<2> **Definition.** *A function μ defined on the sigma-field $\mathcal{A}$ is called a (countably additive, nonnegative) measure if:*

(i) $0 \le \mu A \le \infty$ for each A in $\mathcal{A}$;

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

(iii) if $A_1, A_2, \ldots$ is a countable collection of pairwise disjoint sets in $\mathcal{A}$ then $\mu\left(\cup_i A_i\right) = \sum_i \mu A_i$.

A measure μ for which $\mu\mathcal{X} = 1$ is called a ***probability measure***, and the corresponding $(\mathcal{X}, \mathcal{A}, \mu)$ is called a ***probability space***. For this special case it is traditional to use a symbol like $\mathbb{P}$ for the measure, a symbol like Ω for the set, and a symbol like $\mathcal{F}$ for the sigma-field. A triple $(\Omega, \mathcal{F}, \mathbb{P})$ will always denote a probability space.

Usually the qualifications "countably additive, nonnegative" are omitted, on the grounds that these properties are the most commonly assumed—the most common cases deserve the shortest names. Only when there is some doubt about whether the measures are assumed to have all the properties of Definition <2> should the qualifiers be attached. For example, one speaks of "finitely additive measures" when an analog of property (iii) is assumed only for finite disjoint collections, or "signed measures" when the value of μA is not necessarily nonnegative. When finitely additive or signed measures are under discussion it makes sense to mention explicitly when a particular measure is nonnegative or countably additive, but, in general, you should go with the shorter name.

Where do measures come from? The most basic constructions start from set functions μ defined on small collections of subsets $\mathcal{E}$, such as the collection of all subintervals of the real line. One checks that μ has properties consistent with the requirements of Definition <2>. One seeks to extend the domain of definition while preserving the countable additivity properties of the set function. As you saw in Chapter 1, Theorems guaranteeing existence of such extensions were the culmination of a long sequence of refinements in the concept of integration (Hawkins 1979). They represent one of the great achievements of modern mathematics, even though those theorems now occupy only a handful of pages in most measure theory texts.

Finite additivity has several appealing interpretations (such as the fair-prices of Section 1.5) that have given it ready acceptance as an axiom for a model of real-world uncertainty. Countable additivity is sometimes regarded with suspicion, or justified as a matter of mathematical convenience. (However, see Problem [6] for an equivalent form of countable additivity, which has some claim to intuitive appeal.) It is difficult to develop a simple probability theory without countable additivity, which gives one the licence (for only a small fee) to integrate series term-by-term, differentiate under integrals, and interchange other limiting operations.

The classical constructions are significant for my exposition mostly because they ensure existence of the measures needed to express the basic results of probability theory. I will relegate the details to the Problems and to Appendix A. If you crave a more systematic treatment you might consult one of the many excellent texts on measure theory, such as Royden (1968).

The constructions do not—indeed cannot, in general—lead to countably additive measures on the class of all subsets of a given $\mathcal{X}$. Typically, they extend a set function defined on a class of sets $\mathcal{E}$ to a measure defined on the ***sigma-field*** $\sigma(\mathcal{E})$ ***generated by*** $\mathcal{E}$, or to only slightly larger sigma-fields. By definition,

$$\begin{aligned}\sigma(\mathcal{E}) &:= \text{smallest sigma-field on } \mathcal{X} \text{ containing all sets from } \mathcal{E}\\ &= \{A \subseteq \mathcal{X} : A \in \mathcal{F} \text{ for every sigma-field } \mathcal{F} \text{ with } \mathcal{E} \subseteq \mathcal{F}\}.\end{aligned}$$

The representation given by the second line ensures existence of a smallest sigma-field containing $\mathcal{E}$. The method of definition is analogous to many definitions of "smallest . . . containing a fixed class" in mathematics—think of generated subgroups or linear subspaces spanned by a collection of vectors, for example. For the definition to work one needs to check that sigma-fields have two properties:

(i) If $\{\mathcal{F}_i : i \in \mathcal{I}\}$ is a nonempty collection of sigma-fields on $\mathcal{X}$ then $\cap_{i\in\mathcal{I}}\mathcal{F}_i$, the collection of all the subsets of $\mathcal{X}$ that belong to every $\mathcal{F}_i$, is also a sigma-field.

(ii) For each $\mathcal{E}$ there exists at least one sigma-field $\mathcal{F}$ containing all the sets in $\mathcal{E}$.

You should check property (i) as an exercise. Property (ii) is trivial, because the collection of all subsets of $\mathcal{X}$ is a sigma-field.

REMARK. Proofs of existence of nonmeasurable sets typically depend on some deep set-theoretic principle, such as the Axiom of Choice. Mathematicians who can live with different rules for set theory can have bigger sigma-fields. See Dudley (1989, Section 3.4) or Oxtoby (1971, Section 5) for details.

<3> **Exercise.** Suppose $\mathcal{X}$ consists of five points a, b, c, d, and e. Suppose $\mathcal{E}$ consists of two sets, $E_1 = \{a, b, c\}$ and $E_2 = \{c, d, e\}$. Find the sigma-field generated by $\mathcal{E}$.

SOLUTION: For this simple example we can proceed by mechanical application of the properties that a sigma-field $\sigma(\mathcal{E})$ must possess. In addition to the obvious $\emptyset$ and $\mathcal{X}$, it must contain each of the sets

$$\begin{aligned}F_1 &:= \{a, b\} = E_1 \cap E_2^c \qquad \text{and} \qquad F_2 := \{c\} = E_1 \cap E_2,\\ F_3 &:= \{d, e\} = E_1^c \cap E_2 \qquad \text{and} \qquad F_4 := \{a, b, d, e\} = F_1 \cup F_3.\end{aligned}$$

Further experimentation creates no new members of $\sigma(\mathcal{E})$; the sigma-field consists of the sets

$$\emptyset, \quad F_1, \quad F_2, \quad F_3, \quad F_1 \cup F_3, \quad F_1 \cup F_2 = E_1, \quad F_2 \cup F_3 = E_2, \quad \mathcal{X}.$$

The sets F_1, F_2, F_3 are the ***atoms*** of the sigma-field; every member of $\sigma(\mathcal{E})$ is a union of some collection (possibly empty) of F_i. The only measurable subsets of F_i are the empty set and F_i itself. There are no measurable protons or neutrons hiding inside these atoms. □

An unsystematic construction might work for finite sets, but it cannot generate all members of a sigma-field in general. Indeed, we cannot even hope to list all the members of an infinite sigma-field. Instead we must find a less explicit way to characterize its sets.

<4> **Example.** By definition, the Borel sigma-field on the real line, denoted by $\mathcal{B}(\mathbb{R})$, is the sigma-field generated by the open subsets. We could also denote it by $\sigma(\mathcal{G})$ where $\mathcal{G}$ stands for the class of all open subsets of $\mathbb{R}$. There are several other generating classes for $\mathcal{B}(\mathbb{R})$. For example, as you will soon see, the class $\mathcal{E}$ of all intervals $(-\infty, t]$, with $t \in \mathbb{R}$, is a generating class.

It might appear a hopeless task to prove that $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$ if we cannot explicitly list the members of both sigma-fields, but actually the proof is quite routine. You should try to understand the style of argument because it is often used in probability theory.

The equality of sigma-fields is established by two inclusions, $\sigma(\mathcal{E}) \subseteq \sigma(\mathcal{G})$ and $\sigma(\mathcal{G}) \subseteq \sigma(\mathcal{E})$, both of which follow from more easily established results. First we must prove that $\mathcal{E} \subseteq \sigma(\mathcal{G})$, showing that $\sigma(\mathcal{G})$ is one of the sigma-fields $\mathcal{F}$ that enter into the intersection defining $\sigma(\mathcal{E})$, and hence $\sigma(\mathcal{E}) \subseteq \sigma(\mathcal{G})$. The other inclusion follows similarly if we show that $\mathcal{G} \subseteq \sigma(\mathcal{E})$.

Each interval $(-\infty, t]$ in $\mathcal{E}$ has a representation $\bigcap_{n=1}^{\infty}(-\infty, t+n^{-1})$, a countable intersection of open sets. The sigma-field $\sigma(\mathcal{G})$ contains all open sets, and it is stable under countable intersections. It therefore contains each $(-\infty, t]$. That is, $\mathcal{E} \subseteq \sigma(\mathcal{G})$.

The argument for $\mathcal{G} \subseteq \sigma(\mathcal{E})$ is only slightly harder. It depends on the fact that an open subset of the real line can be written as a countable union of open intervals. Such an interval has a representation $(a, b) = (-\infty, b) \cap (-\infty, a]^c$, and $(-\infty, b) = \bigcup_{n=1}^{\infty}(-\infty, b - n^{-1}]$. That is, every open set can be built up from sets in $\mathcal{E}$ using operations that are guaranteed not to take us outside the sigma-field $\sigma(\mathcal{E})$.

My explanation has been moderately detailed. In a published paper the reasoning would probably be abbreviated to something like "a generating class argument shows that ...," with the routine details left to the reader. □

REMARK. The generating class argument often reduces to an assertion like: $\mathcal{A}$ is a sigma-field and $\mathcal{A} \supseteq \mathcal{E}$, therefore $\mathcal{A} = \sigma(\mathcal{A}) \supseteq \sigma(\mathcal{E})$.

<5> **Example.** A class $\mathcal{E}$ of subsets of a set $\mathcal{X}$ is called a ***field*** if it contains the empty set and is stable under complements, finite unions, and finite intersections. For a field $\mathcal{E}$, write $\mathcal{E}_\delta$ for the class of all possible intersections of countable subclasses of $\mathcal{E}$, and $\mathcal{E}_\sigma$ for the class of all possible unions of countable subclasses of $\mathcal{E}$.

Of course if $\mathcal{E}$ is a sigma-field then $\mathcal{E} = \mathcal{E}_\delta = \mathcal{E}_\sigma$, but, in general, the inclusions $\sigma(\mathcal{E}) \supseteq \mathcal{E}_\delta \supseteq \mathcal{E}$ and $\sigma(\mathcal{E}) \supseteq \mathcal{E}_\sigma \supseteq \mathcal{E}$ will be proper. For example, if $\mathcal{X} = \mathbb{R}$ and $\mathcal{E}$ consists of all finite unions of half open intervals $(a, b]$, with possibly $a = -\infty$ or $b = +\infty$, then the set of rationals does not belong to $\mathcal{E}_\sigma$ and the complement of the same set does not belong to $\mathcal{E}_\delta$.

Let μ be a finite measure on $\sigma(\mathcal{E})$. Even though $\sigma(\mathcal{E})$ might be much larger than either $\mathcal{E}_\sigma$ or $\mathcal{E}_\delta$, a generating class argument will show that all sets in $\sigma(\mathcal{E})$ can be ***inner approximated by*** $\mathcal{E}_\delta$, in the sense that,

$$\mu A = \sup\{\mu F : A \supseteq F \in \mathcal{E}_\delta\} \qquad \text{for each } A \text{ in } \sigma(\mathcal{E}),$$

and ***outer approximated by*** $\mathcal{E}_\sigma$, in the sense that,

$$\mu A = \inf\{\mu G : A \subseteq G \in \mathcal{E}_\sigma\} \qquad \text{for each } A \text{ in } \sigma(\mathcal{E}).$$

REMARK. Incidentally, I chose the letters G and F to remind myself of open and closed sets, which have similar approximation properties for Borel measures on metric spaces—see Problem [12].

It helps to work on both approximation properties at the same time. Denote by $\mathcal{B}_0$ the class of all sets in $\sigma(\mathcal{E})$ that can be both innner and outer approximated. A set B belongs to $\mathcal{B}_0$ if and only if, to each $\epsilon > 0$ there exist $F \in \mathcal{E}_\delta$ and $G \in \mathcal{E}_\sigma$ such that $F \subseteq B \subseteq G$ and $\mu(G \backslash F) < \epsilon$. I'll call the sets F and G an ϵ-sandwich for B.

Trivially $\mathcal{B}_0 \supseteq \mathcal{E}$, because each member of $\mathcal{E}$ belongs to both $\mathcal{E}_\sigma$ and $\mathcal{E}_\delta$. The approximation result will follow if we show that $\mathcal{B}_0$ is a sigma-field, for then we will have $\mathcal{B}_0 = \sigma(\mathcal{B}_0) \supseteq \sigma(\mathcal{E})$.

Symmetry of the definition ensures that $\mathcal{B}_0$ is stable under complements: if $F \subseteq B \subseteq G$ is an ϵ-sandwich for B, then $G^c \subseteq B^c \subseteq F^c$ is an ϵ-sandwich for B^c. To show that $\mathcal{B}_0$ is stable under countable unions, consider a countable collection $\{B_n : n \in \mathbb{N}\}$ of sets from $\mathcal{B}_0$. We need to slice the bread thinner as n gets larger: choose $\epsilon/2^n$-sandwiches $F_n \subseteq B_n \subseteq G_n$ for each n. The union $\cup_n B_n$ is sandwiched between the sets $G := \cup_n G_n$ and $H = \cup_n F_n$; and the sets are close in μ measure because

$$\mu\left(\cup_n G_n \backslash \cup_n F_n\right) \le \sum_n \mu(G_n \backslash F_n) < \sum_n \epsilon/2^n = \epsilon.$$

REMARK. Can you prove this inequality? Do you see why $\cup_n G_n \backslash \cup_n F_n \subseteq \cup_n (G_n \backslash F_n)$ and why countable additivity implies that the measure of a countable union of (not necessarily disjoint) sets is smaller than the sum of their measures? If not, just wait until Section 3, after which you can argue that $\cup_n G_n \backslash \cup_n F_n \le \sum_n (G_n \backslash F_n)$, as an inequality between indicator functions, and $\mu\left(\sum_n (G_n \backslash F_n)\right) = \sum_n \mu(G_n \backslash F_n)$ by Monotone Convergence.

We have an ϵ-sandwich, but the bread might not be of the right type. It is certainly true that $G \in \mathcal{E}_\sigma$ (a countable union of countable unions is a countable union), but the set H need not belong to $\mathcal{E}_\delta$. However, the sets $H_N := \cup_{n \le N} F_n$ do belong to $\mathcal{E}_\delta$, and countable additivity implies that $\mu H_N \uparrow \mu H$.

REMARK. Do you see why? If not, wait for Monotone Convergence again.

□ If we choose a large enough N we have a 2ϵ-sandwich $H_N \subseteq \cup_n B_n \subseteq G$.

The measure $\mathfrak{m}$ on $\mathcal{B}(\mathbb{R})$ for which $\mathfrak{m}(a,b] = b-a$ is called ***Lebesgue measure***. Another sort of generating class argument (see Section 10) can be used to show that the values $\mathfrak{m}(B)$ for B in $\mathcal{B}(\mathbb{R})$ are uniquely determined by the values given to intervals; there can exist at most one measure on $\mathcal{B}(\mathbb{R})$ with the stated property. It is harder to show that at least one such measure exists. Despite any intuitions you might have about length, the construction of Lebesgue measure is not trivial—see Appendix A. Indeed, Henri Lebesgue became famous for proving existence of the measure and showing how much could be done with the new integration theory.

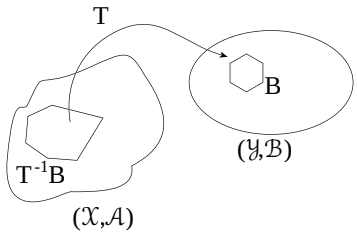
The name Lebesgue measure is also given to an extension of $\mathfrak{m}$ to a measure on a sigma-field, sometimes called the Lebesgue sigma-field, which is slightly larger than $\mathcal{B}(\mathbb{R})$. I will have more to say about the extension in Section 6.

Borel sigma-fields are defined in similar fashion for any topological space $\mathcal{X}$. That is, $\mathcal{B}(\mathcal{X})$ denotes the sigma-field generated by the open subsets of $\mathcal{X}$.

Sets in a sigma-field $\mathcal{A}$ are said to be ***measurable*** or $\mathcal{A}$-measurable. In probability theory they are also called ***events***. Good functions will also be given the title measurable. Try not to get confused when you really need to know whether an object is a set or a function.

2. Measurable functions

Let $\mathcal{X}$ be a set equipped with a sigma-field $\mathcal{A}$, and $\mathcal{Y}$ be a set equipped with a sigma-field $\mathcal{B}$, and T be a function (also called a map) from $\mathcal{X}$ to $\mathcal{Y}$. We say that T ***is*** $\mathcal{A}\backslash\mathcal{B}$***-measurable*** if the inverse image $\{x \in \mathcal{X} : Tx \in B\}$ belongs to $\mathcal{A}$ for each B in $\mathcal{B}$. Sometimes the inverse image is denoted by $\{T \in B\}$ or $T^{-1}B$. Don't be fooled by the T^{-1} notation into treating T^{-1} as a function from $\mathcal{Y}$ into $\mathcal{X}$: it's not, unless T is one-to-one (and onto, if you want to have domain $\mathcal{Y}$). Sometimes an $\mathcal{A}\backslash\mathcal{B}$-measurable map is referred to in abbreviated form as just $\mathcal{A}$-measurable, or just $\mathcal{B}$-measurable, or just measurable, if there is no ambiguity about the unspecified sigma-fields.



For example, if $\mathcal{Y} = \mathbb{R}$ and $\mathcal{B}$ equals the Borel sigma-field $\mathcal{B}(\mathbb{R})$, it is common to drop the $\mathcal{B}(\mathbb{R})$ specification and refer to the map as being $\mathcal{A}$-measurable, or as being Borel measurable if $\mathcal{A}$ is understood and there is any doubt about which sigma-field to use for the real line. *In this book, you may assume that any sigma-field on $\mathbb{R}$ is its Borel sigma-field, unless explicitly specified otherwise.* It can get confusing if you misinterpret where the unspecified sigma-fields live. My advice would be that you imagine a picture showing the two spaces involved, with any missing sigma-field labels filled in.

Sometimes the functions come first, and the sigma-fields are chosen specifically to make those functions measurable.

<6> **Definition.** *Let $\mathcal{H}$ be a class of functions on a set $\mathcal{X}$. Suppose the typical h in $\mathcal{H}$ maps $\mathcal{X}$ into a space $\mathcal{Y}_h$ equipped with a sigma-field $\mathcal{B}_h$. Then the sigma-field $\sigma(\mathcal{H})$ generated by $\mathcal{H}$ is defined as $\sigma\{h^{-1}(B) : B \in \mathcal{B}_h, h \in \mathcal{H}\}$. It is the smallest sigma-field $\mathcal{A}_0$ on $\mathcal{X}$ for which each h in $\mathcal{H}$ is $\mathcal{A}_0\backslash\mathcal{B}_h$-measurable.*

<7> **Example.** If $\mathcal{B} = \sigma(\mathcal{E})$ for some class $\mathcal{E}$ of subsets of $\mathcal{Y}$ then a map T is $\mathcal{A}\backslash\sigma(\mathcal{E})$-measurable if and only if $T^{-1}E \in \mathcal{A}$ for every E in $\mathcal{E}$. You should prove this assertion by checking that $\{B \in \mathcal{B} : T^{-1}B \in \mathcal{A}\}$ is a sigma-field, and then arguing from the definition of a generating class.

In particular, to establish $\mathcal{A}\backslash\mathcal{B}(\mathbb{R})$-measurability of a map into the real line it is enough to check the inverse images of intervals of the form (t, ∞), with t ranging over $\mathbb{R}$. (In fact, we could restrict t to a countable dense subset of $\mathbb{R}$, such as the set of rationals: How would you build an interval (t, ∞) from intervals (t_i, ∞) with rational t_i?) That is, a real-valued function f is Borel-measurable if $\{x \in \mathcal{X} : f(x) > t\} \in \mathcal{A}$ for each real t. There are many similar assertions obtained by using other generating classes for $\mathcal{B}(\mathbb{R})$. Some authors use particular generating classes for the definition of measurability, and then derive facts about inverse images of Borel sets as theorems. □

It will be convenient to consider not just real-valued functions on a set $\mathcal{X}$, but also functions from $\mathcal{X}$ into the extended real line $\overline{\mathbb{R}} := [-\infty, \infty]$. The Borel sigma-field $\mathcal{B}(\overline{\mathbb{R}})$ is generated by the class of open sets, or, more explicitly, by all sets in $\mathcal{B}(\mathbb{R})$ together with the two singletons $\{-\infty\}$ and $\{\infty\}$. It is an easy exercise to show that $\mathcal{B}(\overline{\mathbb{R}})$ is generated by the class of all sets of the form $(t, \infty]$, for t in $\mathbb{R}$, and by the class of all sets of the form $[-\infty, t)$, for t in $\mathbb{R}$. We could even restrict t to any countable dense subset of $\mathbb{R}$.

<8> **Example.** Let a set $\mathcal{X}$ be equipped with a sigma-field $\mathcal{A}$. Let $\{f_n : n \in \mathbb{N}\}$ be a sequence of $\mathcal{A}\backslash\mathcal{B}(\mathbb{R})$-measurable functions from $\mathcal{X}$ into $\mathbb{R}$. Define functions f and g by taking pointwise suprema and infima: $f(x) := \sup_n f_n(x)$ and $g(x) := \inf_n f_n(x)$. Notice that f might take the value $+\infty$, and g might take the value $-\infty$, at some points of $\mathcal{X}$. We may consider both as maps from $\mathcal{X}$ into $\overline{\mathbb{R}}$. (In fact, the whole argument is unchanged if the f_n functions themselves are also allowed to take infinite values.)

The function f is $\mathcal{A}\backslash\mathcal{B}(\overline{\mathbb{R}})$-measurable because

$$\{x : f(x) > t\} = \cup_n \{x : f_n(x) > t\} \in \mathcal{A} \qquad \text{for each real } t :$$

for each fixed x, the supremum of the real numbers $f_n(x)$ is strictly greater than t if and only if $f_n(x) > t$ for at least one n. Example <7> shows why we have only to check inverse images for such intervals.

The same generating class is not as convenient for proving measurability of g. It is not true that an infimum of a sequence of real numbers is strictly greater than t if and only if all of the numbers are strictly greater than t: think of the sequence $\{n^{-1} : n = 1, 2, 3, \ldots\}$, whose infimum is zero. Instead you should argue via the identity $\{x : g(x) < t\} = \cup_n \{x : f_n(x) < t\} \in \mathcal{A}$ for each real t. □

From Example <8> and the representations $\limsup f_n(x) = \inf_{n\in\mathbb{N}} \sup_{m\geq n} f_m(x)$ and $\liminf f_n(x) = \sup_{n\in\mathbb{N}} \inf_{m\geq n} f_m(x)$, it follows that the lim sup or lim inf of a sequence of measurable (real- or extended real-valued) functions is also measurable. In particular, if the limit exists it is measurable.

Measurability is also preserved by the usual algebraic operations—sums, differences, products, and so on—provided we take care to avoid illegal pointwise calculations such as $\infty - \infty$ or $0/0$. There are several ways to establish these stability properties. One of the more direct methods depends on the fact that $\mathbb{R}$ has a countable dense subset, as illustrated by the following argument for sums.

<9> **Example.** Let f and g be $\mathcal{B}(\mathbb{R})$-measurable functions, with pointwise sum $h(x) = f(x) + g(x)$. (I exclude infinite values because I don't want to get caught up with inconclusive discussions of how we might proceed at points x where $f(x) = +\infty$ and $g(x) = -\infty$, or $f(x) = -\infty$ and $g(x) = +\infty$.) How can we prove that h is also a $\mathcal{B}(\mathbb{R})$-measurable function?

It is true that

$$\{x : h(x) > t\} = \cup_{s\in\mathbb{R}} \left(\{x : f(x) = s\} \cap \{x : g(x) > t - s\}\right),$$

and it is true that the set $\{x : f(x) = s\} \cap \{x : g(x) > t - s\}$ is measurable for each s and t, but sigma-fields are not required to have any particular stability properties for uncountable unions. Instead we should argue that at each x for which $f(x)+g(x) > t$ there exists a rational number r such that $f(x) > r > t - g(x)$. Conversely if there is an r lying strictly between $f(x)$ and $t - g(x)$ then $f(x) + g(x) > t$. Thus

$$\{x : h(x) > t\} = \cup_{r\in\mathbb{Q}} \left(\{x : f(x) > r\} \cap \{x : g(x) > t - r\}\right),$$

where $\mathbb{Q}$ denotes the countable set of rational numbers. A countable union of intersections of pairs of measurable sets is measurable. The sum is a measurable function. □

As a little exercise you might try to extend the argument from the last Example to the case where f and g are allowed to take the value $+\infty$ (but not the value $-\infty$). If you want practice at playing with rationals, try to prove measurability of products (be careful with inequalities if dividing by negative numbers) or try Problem [4], which shows why a direct attack on the lim sup requires careful handling of inequalities in the limit.

The real significance of measurability becomes apparent when one works through the construction of integrals with respect to measures, as in Section 4. For the moment it is important only that you understand that the family of all measurable functions is stable under most of the familiar operations of analysis.

<10> **Definition.** *The class $\mathcal{M}(\mathcal{X}, \mathcal{A})$, or $\mathcal{M}(\mathcal{X})$ or just $\mathcal{M}$ for short, consists of all $\mathcal{A}\backslash\mathcal{B}(\overline{\mathbb{R}})$-measurable functions from $\mathcal{X}$ into $\overline{\mathbb{R}}$. The class $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$, or $\mathcal{M}^+(\mathcal{X})$ or just $\mathcal{M}^+$ for short, consists of the nonnegative functions in $\mathcal{M}(\mathcal{X}, \mathcal{A})$.*

If you desired exquisite precision you could write $\mathcal{M}(\mathcal{X}, \mathcal{A}, \overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, to eliminate all ambiguity about domain, range, and sigma-fields.

The collection $\mathcal{M}^+$ is a cone (stable under sums and multiplication of functions by positive constants). It is also stable under products, pointwise limits of sequences,

and suprema or infima of countable collections of functions. It is not a vector space, because it is not stable under subtraction; but it does have the property that if f and g belong to $\mathcal{M}^+$ and g takes only real values, then the positive part $(f-g)^+$, defined by taking the pointwise maximum of $f(x)-g(x)$ with 0, also belongs to $\mathcal{M}^+$. You could adapt the argument from Example <9> to establish the last fact.

It proves convenient to work with $\mathcal{M}^+$ rather than with the whole of $\mathcal{M}$, thereby eliminating many problems with $\infty-\infty$. As you will soon learn, integrals have some convenient properties when restricted to nonnegative functions.

For our purposes, one of the most important facts about $\mathcal{M}^+$ will be the possibility of approximation by ***simple functions*** that is by measurable functions of the form $s := \sum_i \alpha_i A_i$, for finite collections of real numbers α_i and events A_i from $\mathcal{A}$. If the A_i are disjoint, $s(x)$ equals α_i when $x \in A_i$, for some i, and is zero otherwise. If the A_i are not disjoint, the nonzero values taken by s are sums of various subsets of the $\{\alpha_i\}$. Don't forget: the symbol A_i gets interpreted as an indicator function when we start doing algebra. I will write $\mathcal{M}^+_{\text{simple}}$ for the cone of all simple functions in $\mathcal{M}^+$.

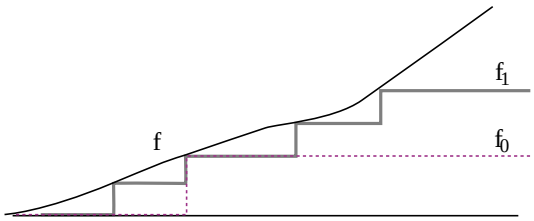
<11> **Lemma.** *For each f in $\mathcal{M}^+$ the sequence $\{f_n\} \subseteq \mathcal{M}^+_{\text{simple}}$, defined by*

$$f_n := 2^{-n} \sum_{i=1}^{4^n} \{f \geq i/2^n\},$$

has the property $0 \leq f_1(x) \leq f_2(x) \leq \ldots \leq f_n(x) \uparrow f(x)$ at every x.

REMARK. The definition of f_n involves algebra, so you must interpret $\{f \geq i/2^n\}$ as the indicator function of the set of all points x for which $f(x) \geq i/2^n$.

Proof. At each x, count the number of nonzero indicator values. If $f(x) \geq 2^n$, all 4^n summands contribute a 1, giving $f_n(x) = 2^n$. If $k2^{-n} \leq f(x) < (k+1)2^{-n}$, for some integer k from $\{0, 1, 2, \ldots, 4^n - 1\}$, then exactly k of the summands contribute a 1, giving $f_n(x) = k2^{-n}$. (Check that the last assertion makes sense when k equals 0.) That is, for $0 \leq f(x) < 2^n$, the function f_n rounds down to an integer multiple of 2^{-n}, from which the convergence and monotone increasing properties follow.



If you do not find the monotonicity assertion convincing, you could argue, more formally, that

$$f_n = \frac{1}{2^{n+1}} \sum_{i=1}^{4^n} 2\left\{f \geq \frac{2i}{2^{n+1}}\right\} \leq \frac{1}{2^{n+1}} \sum_{i=1}^{4\times 4^n} \left(\left\{f \geq \frac{2i}{2^{n+1}}\right\} + \left\{f \geq \frac{2i-1}{2^{n+1}}\right\}\right) = f_{n+1},$$

which reflects the effect of doubling the maximum value and halving the step size when going from the nth to the (n+1)st approximation. □

As an exercise you might prove that the product of functions in $\mathcal{M}^+$ also belongs to $\mathcal{M}^+$, by expressing the product as a pointwise limit of products of simple functions. Notice how the convention $0 \times \infty = 0$ is needed to ensure the correct limit behavior at points where one of the factors is zero.

3. Integrals

Just as $\int_a^b f(x)\,dx$ represents a sort of limiting sum of $f(x)$ values weighted by small lengths of intervals—the $\int$ sign is a long "S", for sum, and the dx is a sort of limiting increment—so can the general integral $\int f(x)\,\mu(dx)$ be defined as a limit of weighted sums but with weights provided by the measure μ. The formal definition involves limiting operations that depend on the assumed measurability of the function f. You can skip the details of the construction (Section 4) by taking the following result as an axiomatic property of the integral.

<12> **Theorem.** *For each measure μ on $(\mathcal{X}, \mathcal{A})$ there is a uniquely determined functional, a map $\tilde{\mu}$ from $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ into $[0, \infty]$, having the following properties:*

(i) $\tilde{\mu}(\mathbb{I}_A) = \mu A$ for each A in $\mathcal{A}$;

(ii) $\tilde{\mu}(0) = 0$, where the first zero stands for the zero function;

(iii) for nonnegative real numbers α, β and functions f, g in $\mathcal{M}^+$,

$$\tilde{\mu}(\alpha f + \beta g) = \alpha \tilde{\mu}(f) + \beta \tilde{\mu}(g);$$

(iv) if f, g are in $\mathcal{M}^+$ and $f \le g$ everywhere then $\tilde{\mu}(f) \le \tilde{\mu}(g)$;

(v) if $f_1, f_2, \ldots$ is a sequence in $\mathcal{M}^+$ with $0 \le f_1(x) \le f_2(x) \le \ldots \uparrow f(x)$ for each x in $\mathcal{X}$ then $\tilde{\mu}(f_n) \uparrow \tilde{\mu}(f)$.

I will refer to (iii) as ***linearity***, even though $\mathcal{M}^+$ is not a vector space. It will imply a linearity property when $\tilde{\mu}$ is extended to a vector subspace of $\mathcal{M}$. Property (iv) is redundant because it follows from (ii) and nonnegativity. Property (ii) is also redundant: put $A = \emptyset$ in (i); or, interpreting $0 \times \infty$ as 0, put $\alpha = \beta = 0$ and $f = g = 0$ in (iii). We need to make sure the bad case $\tilde{\mu} f = \infty$, for all f in $\mathcal{M}^+$, does not slip through if we start stripping away redundant requirements.

Notice that the limit function f in (v) automatically belongs to $\mathcal{M}^+$. The limit assertion itself is called the ***Monotone Convergence property***. It corresponds directly to countable additivity of the measure. Indeed, if $\{A_i : i \in \mathbb{N}\}$ is a countable collection of disjoint sets from $\mathcal{A}$ then the functions $f_n := A_1 + \ldots + A_n$ increase pointwise to the indicator function of $A = \cup_{i \in \mathbb{N}} A_i$, so that Monotone Convergence and linearity imply $\mu A = \sum_i \mu A_i$.

REMARK. You should ponder the role played by $+\infty$ in Theorem <12>. For example, what does $\alpha \tilde{\mu}(f)$ mean if $\alpha = 0$ and $\tilde{\mu}(f) = \infty$? The interpretation depends on the convention that $0 \times \infty = 0$.

In general you should be suspicious of any convention involving $\pm\infty$. Pay careful attention to cases where it operates. For example, how would the five assertions be affected if we adopted a new convention, whereby $0 \times \infty = 6$? Would the Theorem still hold? Where exactly would it fail? I feel uneasy if it is not clear how a convention is disposing of awkward cases. My advice: be very, very

careful with any calculations involving infinity. Subtle errors are easy to miss when concealed within a convention.

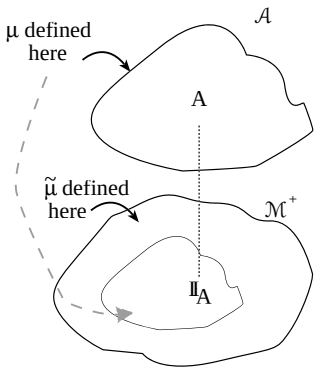
There is a companion to Theorem <12> that shows why it is largely a matter of taste whether one starts from measures or integrals as the more primitive measure theoretic concept.

<13> **Theorem.** *Let $\tilde{\mu}$ be a map from $\mathcal{M}^+$ to $[0,\infty]$ that satisfies properties (ii) through (v) of Theorem <12>. Then the set function defined on the sigma-field $\mathcal{A}$ by (i) is a (countably additive, nonnegative) measure, with $\tilde{\mu}$ the functional that it generates.*

Lemma <11> provides the link between the measure μ and the functional $\tilde{\mu}$. For a given f in $\mathcal{M}^+$, let $\{f_n\}$ be the sequence defined by the Lemma. Then

$$\tilde{\mu} f = \lim_{n\to\infty} \tilde{\mu} f_n = \lim_{n\to\infty} 2^{-n} \sum_{i=1}^{4^n} \mu\{f \ge i/2^n\},$$

the first equality by Monotone Convergence, the second by linearity. The value of $\tilde{\mu} f$ is uniquely determined by μ, as a set function on $\mathcal{A}$. It is even possible to use the equality, or something very similar, as the basis for a direct construction of the integral, from which properties (i) through (v) are then derived, as you will see from Section 4.



In summary: There is a one-to-one correspondence between measures on the sigma-field $\mathcal{A}$ and increasing linear functionals on $\mathcal{M}^+(\mathcal{A})$ with the Monotone Convergence property. To each measure μ there is a uniquely determined functional $\tilde{\mu}$ for which $\tilde{\mu}(\mathbb{I}_A) = \mu(A)$ for every A in $\mathcal{A}$. The functional $\tilde{\mu}$ is usually called an ***integral*** with respect to μ, and is variously denoted by $\int f\, d\mu$ or $\int f(x)\, \mu(dx)$ or $\int_{\mathcal{X}} f\, d\mu$ or $\int f(x)\, d\mu(x)$. With the de Finetti notation, where we identify a set A with its indicator function, the functional $\tilde{\mu}$ is just an extension of μ from a smaller domain (indicators of sets in $\mathcal{A}$) to a larger domain (all of $\mathcal{M}^+$). Accordingly, we should have no qualms about denoting it by the same symbol. I will write μf for the integral. With this notation, assertion (i) of Theorem <12> becomes: $\mu A = \mu A$ for all A in $\mathcal{A}$. You probably can't tell that the A on the left-hand side is an indicator function and the μ is an integral, but you don't need to be able to tell—that is precisely what (i) asserts.

REMARK. In elementary algebra we rely on parentheses, or precedence, to make our meaning clear. For example, both $(ax)+b$ and $ax+b$ have the same meaning, because multiplication has higher precedence than addition. With traditional notation, the $\int$ and the $d\mu$ act like parentheses, enclosing the integrand and separating it from following terms. With linear functional notation, we sometimes need explicit parentheses to make the meaning unambiguous. As a way of eliminating some parentheses, I often work with the convention that integration has lower precedence than exponentiation, multiplication, and division, but higher precedence than addition or subtraction. Thus I intend you to read $\mu fg + 6$ as $(\mu(fg)) + 6$. I would write $\mu(fg+6)$ if the 6 were part of the integrand.

Some of the traditional notations also remove ambiguity when functions of several variables appear in the integrand. For example, in $\int f(x,y)\,\mu(dx)$ the y variable is held fixed while the μ operates on the first argument of the function. When a similar ambiguity might arise with linear functional notation, I will append a superscript, as in $\mu^x f(x,y)$, to make clear which variable is involved in the integration.

<14> **Example.** Suppose μ is a finite measure (that is, $\mu\mathcal{X} < \infty$) and f is a function in $\mathcal{M}^+$. Then $\mu f < \infty$ if and only if $\sum_{n=1}^{\infty} \mu\{f \geq n\} < \infty$.

The assertion is just a pointwise inequality in disguise. By considering separately values for which $k \leq f(x) < k+1$, for $k = 0, 1, 2, \ldots$, you can verify the pointwise inequality between functions,

$$\textstyle\sum_{n=1}^{\infty}\{f \geq n\} \leq f \leq 1 + \sum_{n=1}^{\infty}\{f \geq n\}.$$

In fact, the sum on the left-hand side defines $\lfloor f(x) \rfloor$, the largest integer $\leq f(x)$, and the right-hand side denotes the smallest integer $> f(x)$. From the leftmost inequality,

$$\begin{aligned}
\mu f &\geq \mu\textstyle\left(\sum_{n=1}^{\infty}\{f \geq n\}\right) && \text{increasing} \\
&= \lim_{N\to\infty} \mu\left(\sum_{n=1}^{N}\{f \geq n\}\right) && \text{Monotone Convergence} \\
&= \lim_{N\to\infty} \textstyle\sum_{n=1}^{N} \mu\{f \geq n\} && \text{linearity} \\
&= \textstyle\sum_{n=1}^{\infty} \mu\{f \geq n\}.
\end{aligned}$$

A similar argument gives a companion upper bound. Thus the pointwise inequality integrates out to $\sum_{n=1}^{\infty} \mu\{f \geq n\} \leq \mu f \leq \mu\mathcal{X} + \sum_{n=1}^{\infty} \mu\{f \geq n\}$, from which the asserted equivalence follows. □

Extension of the integral to a larger class of functions

Every function f in $\mathcal{M}$ can be decomposed into a difference $f = f^+ - f^-$ of two functions in $\mathcal{M}^+$, where $f^+(x) := \max\big(f(x), 0\big)$ and $f^-(x) := \max\big(-f(x), 0\big)$. To extend μ from $\mathcal{M}^+$ to a linear functional on $\mathcal{M}$ we should define $\mu f := \mu f^+ - \mu f^-$. This definition works if at least one of μf^+ and μf^- is finite; otherwise we get the dreaded $\infty - \infty$. If both $\mu f^+ < \infty$ and $\mu f^- < \infty$ (or equivalently, f is measurable and $\mu|f| < \infty$) the function f is said to be ***integrable*** or μ-integrable. The linearity property (iii) of Theorem <12> carries over partially to $\mathcal{M}$ if $\infty - \infty$ problems are excluded, although it becomes tedious to handle all the awkward cases involving $\pm\infty$. The constants α and β need no longer be nonnegative. Also if both f and g are integrable and if $f \leq g$ then $\mu f \leq \mu g$, with obvious extensions to certain cases involving ∞.

<15> **Definition.** *The set of all real-valued, μ-integrable functions in $\mathcal{M}$ is denoted by $\mathcal{L}^1(\mu)$, or $\mathcal{L}^1(\mathcal{X}, \mathcal{A}, \mu)$.*

The set $\mathcal{L}^1(\mu)$ is a vector space (stable under pointwise addition and multiplication by real numbers). The integral μ defines an increasing linear functional on $\mathcal{L}^1(\mu)$, in the sense that $\mu f \geq \mu g$ if $f \geq g$ pointwise. The Monotone Convergence property implies other powerful limit results for functions in $\mathcal{L}^1(\mu)$, as described in Section 5. By restricting μ to $\mathcal{L}^1(\mu)$, we eliminate problems with $\infty - \infty$.

For each f in $\mathcal{L}^1(\mu)$, its $\mathcal{L}^1$ ***norm*** is defined as $\|f\|_1 := \mu|f|$. Strictly speaking, $\|\cdot\|_1$ is only a seminorm, because $\|f\|_1 = 0$ need not imply that f is the zero function—as you will see in Section 6, it implies only that $\mu\{f \neq 0\} = 0$. It is common practice to ignore the small distinction and refer to $\|\cdot\|_1$ as a norm on $\mathcal{L}^1(\mu)$.

<16> **Example.** Let Ψ be a convex, real-valued function on $\mathbb{R}$. The function Ψ is measurable (because $\{\Psi \le t\}$ is an interval for each real t), and for each x_0 in $\mathbb{R}$ there is a constant α such that $\Psi(x) \ge \Psi(x_0) + \alpha(x - x_0)$ for all x (Appendix C).

Let $\mathbb{P}$ be a probability measure, and X be an integrable random variable. Choose $x_0 := \mathbb{P}X$. From the inequality $\Psi(x) \ge -|\Psi(x_0)| - |\alpha|(|x| + |x_0|)$ we deduce that $\mathbb{P}\Psi(X)^- \le |\Psi(x_0)| + |\alpha|(\mathbb{P}|X| + |x_0|) < \infty$. Thus we should have no $\infty - \infty$ worries in taking expectations (that is, integrating with respect to $\mathbb{P}$) to deduce that $\mathbb{P}\Psi(X) \ge \Psi(\mathbb{P}X) + \alpha(\mathbb{P}X - x_0) = \Psi(\mathbb{P}X)$, a result known as ***Jensen's inequality***. One way to remember the direction of the inequality is to note that $0 \le \operatorname{var}(X) = \mathbb{P}X^2 - (\mathbb{P}X)^2$, which corresponds to the case $\Psi(x) = x^2$. □

Integrals with respect to Lebesgue measure

Lebesgue measure $\mathfrak{m}$ on $\mathcal{B}(\mathbb{R})$ corresponds to length: $\mathfrak{m}[a, b] = b - a$ for each interval. I will occasionally revert to the traditional ways of writing such integrals,

$$\mathfrak{m}f = \int f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx \qquad \text{and} \qquad \mathfrak{m}^x\big(f(x)\{a \le x \le b\}\big) = \int_a^b f(x)dx.$$

Don't worry about confusing the Lebesgue integral with the Riemann integral over finite intervals. Whenever the Riemann is well defined, so is the Lebesgue, and the two sorts of integral have the same value. The Lebesgue is a more general concept. Indeed, facts about the Riemann are often established by an appeal to theorems about the Lebesgue. You do not have to abandon what you already know about integration over finite intervals.

The improper Riemann integral, $\int_{-\infty}^{\infty} f(x)\,dx = \lim_{n\to\infty}\int_{-n}^{n} f(x)\,dx$, also agrees with the Lebesgue integral provided $\mathfrak{m}|f| < \infty$. If $\mathfrak{m}|f| = \infty$, as in the case of the function $f(x) := \sum_{n=1}^{\infty}\{n \le x < n+1\}(-1)^n/n$, the improper Riemann integral might exist as a finite limit, while the Lebesgue integral $\mathfrak{m}f$ does not exist.

*4. Construction of integrals from measures

To construct the integral $\tilde{\mu}$ as a functional on $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$, starting from a measure μ on the sigma-field $\mathcal{A}$, we use approximation from below by means of simple functions.

First we must define $\tilde{\mu}$ on $\mathcal{M}^+_{\text{simple}}$. The representation of a simple function as a linear combination of indicator functions is not unique, but the additivity properties of the measure μ will let us use any representation to define the integral. For example, if $s := 3A_1 + 7A_2 = 3A_1A_2^c + 10A_1A_2 + 7A_1^cA_2$, then

$$3\mu(A_1) + 7\mu(A_2) = 3\mu(A_1A_2^c) + 10\mu(A_1A_2) + 7\mu(A_1^cA_2).$$

More generally, if $s := \sum_i \alpha_i A_i$ has another representation $s = \sum_j \beta_j B_j$, then $\sum_i \alpha_i \mu A_i = \sum_j \beta_j \mu B_j$. Proof? Thus we can uniquely define $\tilde{\mu}(s)$ for a simple function $s := \sum_i \alpha_i A_i$ by $\tilde{\mu}(s) := \sum_i \alpha_i \mu A_i$.

Define the increasing functional $\tilde{\mu}$ on $\mathcal{M}^+$ by

$$\tilde{\mu}(f) := \sup\{\tilde{\mu}(s) : f \geq s \in \mathcal{M}^+_{\text{simple}}\}.$$

That is, the integral of f is a supremum of integrals of nonnegative simple functions less than f.

From the representation of simple functions as linear combinations of disjoint sets in $\mathcal{A}$, it is easy to show that $\tilde{\mu}(\mathbb{I}_A) = \mu A$ for every A in $\mathcal{A}$. It is also easy to show that $\tilde{\mu}(0) = 0$, and $\tilde{\mu}(\alpha f) = \alpha \tilde{\mu}(f)$ for nonnegative real α, and

<17> $$\tilde{\mu}(f+g) \geq \tilde{\mu}(f) + \tilde{\mu}(g).$$

The last inequality, which is usually referred to as the superadditivity property, follows from the fact that if $f \geq u$ and $g \geq v$, and both u and v are simple, then $f + g \geq u + v$ with $u + v$ simple.

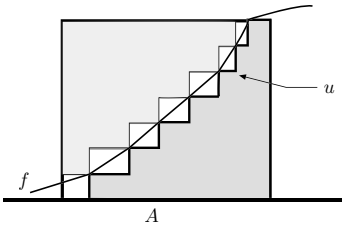
Only the Monotone Convergence property and the companion to <17>,

<18> $$\tilde{\mu}(f+g) \leq \tilde{\mu}(f) + \tilde{\mu}(g),$$

require real work. Here you will see why measurability is needed.

Proof of inequality <18>. Let s be a simple function $\leq f + g$, and let ϵ be a small positive number. It is enough to construct simple functions u, v with $u \leq f$ and $v \leq g$ such that $u + v \geq (1-\epsilon)s$. For then $\tilde{\mu} f + \tilde{\mu} g \geq \tilde{\mu} u + \tilde{\mu} v \geq (1-\epsilon)\tilde{\mu} s$, from which the subadditivity inequality <18> follows by taking a supremum over simple functions then letting ϵ tend to zero.

For simplicity of notation I will assume s to be very simple: $s := A$. You can repeat the argument for each A_i in a representation $\sum_i \alpha_i A_i$ with disjoint A_i to get the general result. Suppose $\epsilon = 1/m$ for some positive integer m. Write ℓ_j for j/m. Define simple functions

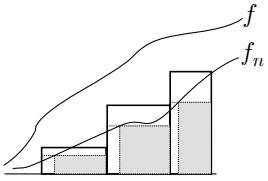


$$u := A\{f \geq 1\} + \sum_{j=1}^m A\left\{\ell_{j-1} \leq f < \ell_j\right\} \ell_{j-1},$$
$$v := \sum_{j=1}^m A\left\{\ell_{j-1} \leq f < \ell_j\right\} \left(1 - \ell_j\right).$$

The measurability of f ensures $\mathcal{A}$-measurability of all the sets entering into the definitions of u and v. For the inequality $v \leq g$, notice that $f + g \geq 1$ on A, so $g > 1 - \ell_j = v$ when $\ell_{j-1} \leq f < \ell_j$ on A. Finally, note that the simple functions were chosen so that

$$u + v = A\{f \geq 1\} + \sum_{j=1}^m A\left\{\ell_{j-1} \leq f < \ell_j\right\}(1-\epsilon) \geq (1-\epsilon)\, A,$$

□ as desired.



Proof of the Monotone Convergence property. Suppose $f_n \in \mathcal{M}^+$ and $f_n \uparrow f$. Suppose $f \geq s := \sum \alpha_i A_i$, with the A_i disjoint sets in $\mathcal{A}$ and $\alpha_i > 0$. Define approximating simple functions $s_n := \sum_i (1-\epsilon)\alpha_i A_i\{f_n \geq (1-\epsilon)\alpha_i\}$. Clearly $s_n \leq f_n$. The

simple function s_n is one of those that enters into the supremum defining $\tilde{\mu} f_n$. It follows that

$$\tilde{\mu} f_n \geq \tilde{\mu}(s_n) = (1-\epsilon) \sum_i \alpha_i \mu \left(A_i \{ f_n \geq (1-\epsilon)\alpha_i \} \right).$$

On the set A_i the functions f_n increase monotonely to f, which is $\geq \alpha_i$. The sets $A_i\{f_n \geq (1-\epsilon)\alpha_i\}$ expand up to the whole of A_i. Countable additivity implies that the μ measures of those sets increase to μA_i. It follows that

$$\liminf \tilde{\mu} f_n \geq \limsup \tilde{\mu} s_n \geq (1-\epsilon)\tilde{\mu} s.$$

□ Take a supremum over simple $s \leq f$ then let ϵ tend to zero to complete the proof.

5. Limit theorems

Theorem <13> identified an integral on $\mathcal{M}^+$ as an increasing linear functional with the Monotone Convergence property :

<19> $$\text{if } 0 \leq f_n \uparrow \text{ then } \mu \left(\lim_{n\to\infty} f_n \right) = \lim_{n\to\infty} \mu f_n.$$

Two direct consequences of this limit property have important applications throughout probability theory. The first, ***Fatou's Lemma***, asserts a weaker limit property for nonnegative functions when the convergence and monotonicity assumptions are dropped. The second, ***Dominated Convergence***, drops the monotonicity and nonnegativity but imposes an extra domination condition on the convergent sequence $\{f_n\}$. I have slowly realized over the years that many simple probabilistic results can be established by Dominated Convergence arguments. The Dominated Convergence Theorem is the Swiss Army Knife of probability theory.

It is important that you understand why some conditions are needed before we can interchange integration (which is a limiting operation) with an explicit limit, as in <19>. Variations on the following example form the basis for many counterexamples.

<20> **Example.** Let μ be Lebesgue measure on $\mathcal{B}[0,1]$ and let $\{\alpha_n\}$ be a sequence of positive numbers. The function $f_n(x) := \alpha_n\{0 < x < 1/n\}$ converges to zero, pointwise, but its integral $\mu(f_n) = \alpha_n/n$ need not converge to zero. For example, $\alpha_n = n^2$ gives $\mu f_n \to \infty$; the integrals diverge. And

$$\alpha_n = \begin{cases} 6n & \text{for } n \text{ even} \\ 3n & \text{for } n \text{ odd} \end{cases} \qquad \text{gives} \qquad \mu f_n = \begin{cases} 6 & \text{for } n \text{ even} \\ 3 & \text{for } n \text{ odd.} \end{cases}$$

□ The integrals oscillate.

<21> **Fatou's Lemma.** *For every sequence $\{f_n\}$ in $\mathcal{M}^+$ (not necessarily convergent), $\mu(\liminf_{n\to\infty} f_n) \leq \liminf_{n\to\infty} \mu(f_n)$.*

Proof. Write f for $\liminf f_n$. Remember what a $\liminf$ means. Define $g_n := \inf_{m\geq n} f_m$. Then $g_n \leq f_n$ for every n and the $\{g_n\}$ sequence increases monotonely to the function f. By Monotone Convergence, $\mu f = \lim_{n\to\infty} \mu g_n$. By the increasing property, $\mu g_n \leq \mu f_n$ for each n, and hence $\lim_{n\to\infty} \mu g_n \leq \liminf_{n\to\infty} \mu f_n$. □

For dominated sequences of functions, a splicing together of two Fatou Lemma assertions gives two lim inf consequences that combine to produce a limit result. (See Problem [10] for a generalization.)

<22> **Dominated Convergence.** *Let $\{f_n\}$ be a sequence of μ-integrable functions for which $\lim_n f_n(x)$ exists for all x. Suppose there exists a μ-integrable function F, which does not depend on n, such that $|f_n(x)| \le F(x)$ for all x and all n. Then the limit function is integrable and $\mu(\lim_{n\to\infty} f_n) = \lim_{n\to\infty} \mu f_n$.*

Proof. The limit function is also bounded in absolute value by F, and hence it is integrable.

Apply Fatou's Lemma to the two sequences $\{F + f_n\}$ and $\{F - f_n\}$ in $\mathcal{M}^+$, to get

$$\mu(\liminf(F + f_n)) \le \liminf \mu(F + f_n) = \liminf (\mu F + \mu f_n),$$
$$\mu(\liminf(F - f_n)) \le \liminf \mu(F - f_n) = \liminf (\mu F - \mu f_n).$$

Simplify, using the fact that a lim inf is the same as a lim for convergent sequences.

$$\mu(F \pm \lim f_n)) \le \mu F + \liminf (\pm \mu f_n).$$

Notice that we cannot yet assert that the lim inf on the right-hand side is actually a limit. The negative sign turns a lim inf into a lim sup.

$$\mu F \pm \mu(\lim f_n) \le \begin{cases} \mu F + \liminf \mu f_n \\ \mu F - \limsup \mu f_n \end{cases}$$

Cancel out the finite number μF then rearrange, leaving

$$\limsup \mu f_n \le \mu(\lim f_n) \le \liminf \mu f_n.$$

□ The convergence assertion follows.

REMARK. You might well object to some of the steps in the proof on $\infty - \infty$ grounds. For example, what does $F(x) + f_n(x)$ mean at a point where $F(x) = \infty$ and $f_n(x) = -\infty$? To eliminate such problems, replace F by $F\{F < \infty\}$ and f_n by $f_n\{F < \infty\}$, then appeal to Lemma <26> in the next Section to ensure that the integrals are not affected.

The function F is said to ***dominate*** the sequence $\{f_n\}$. The assumption in Theorem <22> could also be written as $\mu\left(\sup_n |f_n|\right) < \infty$, with $F := \sup_n |f_n|$ as the dominating function. It is a common mistake amongst students new to the result to allow F to depend on n.

Dominated Convergence turns up in many situations that you might not at first recognize as examples of an interchange in the order of two limit procedures.

<23> **Example.** Do you know why

$$\frac{d}{dt}\int_0^1 e^{xt} x^{5/2}(1-x)^{3/2}\,dx = \int_0^1 e^{xt} x^{7/2}(1-x)^{3/2}\,dx\,?$$

Of course I just differentiated under the integral sign, but why is that allowed? The neatest justification uses a Dominated Convergence argument.

More generally, for each t in an interval $(-\delta, \delta)$ about the origin let $f(\cdot, t)$ be a μ-integrable function on $\mathcal{X}$, such that the function $f(x, \cdot)$ is differentiable in $(-\delta, \delta)$

for each x. We need to justify taking the derivative at $t = 0$ inside the integral, to conclude that

<24> $$\frac{d}{dt}\left(\mu^x f(x,t)\right)\Big|_{t=0} = \mu^x\left(\frac{\partial}{\partial t} f(x,t)\Big|_{t=0}\right).$$

Domination of the partial derivative will suffice.

Write $g(t)$ for $\mu^x f(x,t)$ and $\Delta(x,t)$ for the partial derivative $\frac{\partial}{\partial t} f(x,t)$. Suppose there exists a μ-integrable function M such that

$$|\Delta(x,t)| \le M(x) \qquad \text{for all } x, \text{ all } t \in (-\delta, \delta).$$

To establish <24>, it is enough to show that

<25> $$\frac{g(h_n) - g(0)}{h_n} \to \mu^x \Delta(x,0)$$

for every sequence $\{h_n\}$ of nonzero real numbers tending to zero. (Please make sure that you understand why continuous limits can be replaced by sequential limits in this way. It is a common simplification.) With no loss of generality, suppose $\delta > h_n > 0$ for all n. The ratio on the left-hand side of <25> equals the μ integral of the function $f_n(x) := (f(x,h_n) - f(x,0))/h_n$. By assumption, $f_n(x) \to \Delta(x,0)$ for every x. The sequence $\{f_n\}$ is dominated by M: by the mean-value theorem, for each x there exists a t_x in $(-h_n, h_n) \subseteq (-\delta, \delta)$ for which $|f_n(x)| = |\Delta(x,t_x)| \le M(x)$. An appeal to Dominated Convergence completes the argument. □

6. Negligible sets

A set N in $\mathcal{A}$ for which $\mu N = 0$ is said to be ***μ-negligible***. (Some authors use the term μ-null, but I find it causes confusion with null as a name for the empty set.) As the name suggests, we can usually ignore bad things that happen only on a negligible set. A property that holds everywhere except possibly for those x in a μ-negligible set of points is said to hold ***μ-almost everywhere*** or ***μ-almost surely***, abbreviated to ***a.e.*** $[\mu]$ or ***a.s.*** $[\mu]$, with the $[\mu]$ omitted when understood.

There are several useful facts about negligible sets that are easy to prove and exceedingly useful to have formally stated. They depend on countable additivity, via its Monotone Convergence generalization. I state them only for nonnegative functions, leaving the obvious extensions for $\mathcal{L}^1(\mu)$ to you.

<26> **Lemma.** *For every measure μ:*

(i) if $g \in \mathcal{M}^+$ and $\mu g < \infty$ then $g < \infty$ a.e. $[\mu]$*;*

(ii) if $g, h \in \mathcal{M}^+$ and $g = h$ a.e. $[\mu]$ *then $\mu g = \mu h$;*

(iii) if $N_1, N_2, \ldots$ is a sequence of negligible sets then $\bigcup_i N_i$ is also negligible;

(iv) if $g \in \mathcal{M}^+$ and $\mu g = 0$ then $g = 0$ a.e. $[\mu]$.

Proof. For (i): Integrate out the inequality $g \ge n\{g = \infty\}$ for each positive integer n to get $\infty > \mu g \ge n\mu\{g = \infty\}$. Let n tend to infinity to deduce that $\mu\{g = \infty\} = 0$.

For (ii): Invoke the increasing and Monotone Convergence properties of integrals, starting from the pointwise bound $h \le g + \infty\{h \ne g\} = \lim_n (g + n\{h \ne g\})$

to deduce that $\mu h \le \lim_n (\mu g + n\mu\{h \ne g\}) = \mu g$. Reverse the roles of g and h to get the reverse inequality.

For (iii): Invoke Monotone Convergence for the right-hand side of the pointwise inequality $\cup_i N_i \le \sum_i N_i$ to get $\mu(\cup_i N_i) \le \mu\left(\sum_i N_i\right) = \sum_i \mu N_i = 0$.

For (iv): Put $N_n := \{g \ge 1/n\}$ for $n = 1, 2, \ldots$. Then $\mu N_n \le n\mu g = 0$, from which it follows that $\{g > 0\} = \bigcup_n N_n$ is negligible. □

REMARK. Notice the appeals to countable additivity, via the Monotone Convergence property, in the proofs. Results such as (iv) fail without countable additivity, which might trouble those brave souls who would want to develop a probability theory using only finite additivity.

Property (iii) can be restated as: if $A \in \mathcal{A}$ and A is covered by a countable family of negligible sets then A is negligible. Actually we can drop the assumption that $A \in \mathcal{A}$ if we enlarge the sigma-field slightly.

<27> **Definition.** *The μ-completion of the sigma-field $\mathcal{A}$ is the class $\mathcal{A}_\mu$ of all those sets B for which there exist sets A_0, A_1 in $\mathcal{A}$ with $A_0 \subseteq B \subseteq A_1$ and $\mu(A_1 \backslash A_0) = 0$.*

You should check that $\mathcal{A}_\mu$ is a sigma-field and that μ has a unique extension to a measure on $\mathcal{A}_\mu$ defined by $\mu B := \mu A_0 = \mu A_1$, with A_0 and A_1 as in the Definition. More generally, for each f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A}_\mu)$, you should show that there exist functions f_0, g_0 in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ for which $f_0 \le f \le f_0 + g_0$ and $\mu g_0 = 0$. Of course, we then have $\mu f := \mu f_0$.

The Lebesgue sigma-field on the real line is the completion of the Borel sigma-field with respect to Lebesgue measure.

<28> **Example.** Here is one of the standard methods for proving that some measurable set A has zero μ measure. Find a measurable function f for which $f(x) > 0$, for all x in A, and $\mu(fA) = 0$. From part (iv) of Lemma <26> deduce that $fA = 0$ a.e. $[\mu]$. That is, $f(x) = 0$ for almost all x in A. The set $A = \{x \in A : f(x) > 0\}$ must be negligible. □

Many limit theorems in probability theory assert facts about sequences that hold only almost everywhere.

<29> **Example.** (Generalized Borel-Cantelli lemma) Suppose $\{f_n\}$ is a sequence in $\mathcal{M}^+$ for which $\sum_n \mu f_n < \infty$. By Monotone Convergence, $\mu \sum_n f_n = \sum_n \mu f_n < \infty$. Part (i) of Lemma <26> then gives $\sum_n f_n(x) < \infty$ for μ almost all x.

For the special case of probability measure with each f_n an indicator function of a set in $\mathcal{A}$, the convergence property is called the ***Borel-Cantelli lemma***: If $\sum_n \mathbb{P}A_n < \infty$ then $\sum_n A_n < \infty$ almost surely. That is,

$$\mathbb{P}\{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\} = 0,$$

a trivial result that, nevertheless, is the basis for much probabilistic limit theory. The event in the last display is often written in abbreviated form, $\{A_n \text{ i. o.}\}$.

REMARK. For sequences of independent events, there is a second part to the Borel-Cantelli lemma (Problem [1]), which asserts that if $\sum_n \mathbb{P}A_n = \infty$ then $\mathbb{P}\{A_n \text{ i. o.}\} = 1$. Problem [2] establishes an even stronger converse, replacing independence by a weaker limit property.

The Borel-Cantelli argument often takes the following form when invoked to establish almost sure convergence. You should make sure you understand the method, because the details are usually omitted in the literature.

Suppose $\{X_n\}$ is a sequence of random variables (all defined on the same Ω) for which $\sum_n \mathbb{P}\{|X_n| > \epsilon\} < \infty$ for each $\epsilon > 0$. By Borel-Cantelli, to each $\epsilon > 0$ there is a $\mathbb{P}$-negligible set $N(\epsilon)$ for which $\sum_n \{|X_n(\omega)| > \epsilon\} < \infty$ if $\omega \in N(\epsilon)^c$. A sum of integers converges if and only if the summands are eventually zero. Thus to each ω in $N(\epsilon)^c$ there exists a finite $n(\epsilon, \omega)$ such that $|X_n(\omega)| \le \epsilon$ when $n \ge n(\epsilon, \omega)$.

We have an uncountable family of negligible sets $\{N(\epsilon) : \epsilon > 0\}$. We are allowed to neglect only countable unions of negligible sets. Replace ϵ by a sequence of values such as 1, 1/2, 1/3, 1/4, ..., tending to zero. Define $N := \bigcup_{k=1}^{\infty} N(1/k)$, which, by part (iii) of Lemma <26>, is negligible. For each ω in N^c we have $|X_n(\omega)| \le 1/k$ when $n \ge n(1/k, \omega)$. Consequently, $X_n(\omega) \to 0$ as $n \to \infty$ for each ω in N^c; the sequence $\{X_n\}$ converges to zero almost surely. □

For measure theoretic arguments with a fixed μ, it is natural to treat as identical those functions that are equal almost everywhere. Many theorems have trivial modifications with equalities replaced by almost sure equalities, and convergence replaced by almost sure convergence, and so on. For example, Dominated Convergence holds in a slightly strengthened form:

> Let $\{f_n\}$ be a sequence of measurable functions for which $f_n(x) \to f(x)$ at μ *almost* all x. Suppose there exists a μ-integrable function F, which does not depend on n, such that $|f_n(x)| \le F(x)$ for μ *almost* all x and all n. Then $\mu f_n \to \mu f$.

Most practitioners of probability learn to ignore negligible sets (and then suffer slightly when they come to some stochastic process arguments where the handling of uncountable families of negligible sets requires more delicacy). For example, if I could show that a sequence $\{f_n\}$ converges almost everywhere I would hardly hesitate to write: Define $f := \lim_n f_n$. What happens at those x where $f_n(x)$ does not converge? If hard pressed I might write:

$$\text{Define } f(x) := \begin{cases} \lim_n f_n(x) & \text{on the set where the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

You might then wonder if the function so-defined were measurable (it is), or if the set where the limit exists is measurable (it is). A sneakier solution would be to write: Define $f(x) := \limsup_n f_n(x)$. It doesn't much matter what happens on the negligible set where the limsup is not equal to the liminf, which happens only when the limit does not exist.

A more formal way to equate functions equal almost everywhere is to work with equivalence classes, $[f] := \{g \in \mathcal{M} : f = g \text{ a.e. } [\mu]\}$. The almost sure equivalence also partitions $\mathcal{L}^1(\mu)$ into equivalence classes, for which we can define $\mu[f] := \mu g$ for an arbitrary choice of g from $[f]$. The collection of all these equivalence classes is denoted by $L^1(\mu)$. The L^1 norm, $\|[f]\|_1 := \|f\|_1$, is a true norm on L^1, because $[f]$ equals the equivalence class of the zero function when $\|[f]\|_1 = 0$. Few authors are careful about maintaining the distinction between f and $[f]$, or between $L^1(\mu)$ and $\mathcal{L}^1(\mu)$.

*7. L^p spaces

For each real number p with $p \geq 1$ the $\mathcal{L}^p$-***norm*** is defined on $\mathcal{M}(\mathcal{X}, \mathcal{A}, \mu)$ by $\|f\|_p := (\mu|f|^p)^{1/p}$. Problem [17] shows that the map $f \mapsto \|f\|_p$ satisfies the triangle inequality, $\|f+g\|_p \leq \|f\|_p + \|g\|_p$, at least when restricted to real-valued functions in $\mathcal{M}$.

As with the $\mathcal{L}^1$-norm, it is not quite correct to call $\|\cdot\|_p$ a norm, for two reasons: there are measurable functions for which $\|f\|_p = \infty$, and there are nonzero measurable functions for which $\|f\|_p = 0$. We avoid the first complication by restricting attention to the vector space $\mathcal{L}^p := \mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$ of all real-valued, $\mathcal{A}$-measurable functions for which $\|f\|_p < \infty$. We could avoid the second complication by working with the vector space $L^p := L^p(\mathcal{X}, \mathcal{A}, \mu)$ of μ-equivalence classes of functions in $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$. That is, the members of L^p are the μ-equivalence classes, $[f] := \{g \in \mathcal{L}^p : g = f \text{ a.e. } [\mu]\}$, with f in $\mathcal{L}^p$. (See Problem [20] for the limiting case, $p = \infty$.)

> REMARK. The correct term for $\|\cdot\|_p$ on $\mathcal{L}^p$ is ***pseudonorm***, meaning that it has all the properties of a norm (triangle inequality, and $\|cf\| = |c|\,\|f\|$ for real constants c) except that it might be zero for nonzero functions. Again, few authors are careful about maintaining the distinction between $\mathcal{L}^p$ and L^p.

Problem [19] shows that the norm defines a complete pseudometric on $\mathcal{L}^p$ (and a complete metric on L^p). That is, if $\{f_n\}$ is a Cauchy sequence of functions in $\mathcal{L}^p$ (meaning that $\|f_n - f_m\|_p \to 0$ as $\min(m,n) \to \infty$) then there exists a function f in $\mathcal{L}^p$ for which $\|f_n - f\|_n \to 0$. The limit function f is unique up to a μ-equivalence.

For our purposes, the case where p equals 2 will be the most important. The pseudonorm is then generated by an inner product (or, more correctly, a "pseudo" inner product), $\langle f, g\rangle := \mu(fg)$. That is, $\|f\|_2^2 := \langle f, f\rangle$. The inner product has the properties:

(a) $\langle \alpha f + \beta g, h\rangle = \alpha\langle f, h\rangle + \beta\langle g, h\rangle$ for all real α, β all f, g, h in $\mathcal{L}^2$;

(b) $\langle f, g\rangle = \langle g, f\rangle$ for all f, g in $\mathcal{L}^2$;

(c) $\langle f, f\rangle \geq 0$ with equality if and only if $f = 0$ a.e. $[\mu]$.

If we work with the equivalence classes of L^2 then (c) is replaced by the assertion that $\langle [f], [f]\rangle$ equals zero if and only if $[f]$ is zero, as required for a true inner product.

A vector space equipped with an inner product whose corresponding norm defines a complete metric is called a ***Hilbert space***, a generalization of ordinary Euclidean space. Arguments involving Hilbert spaces look similar to their analogs for Euclidean space, with an occasional precaution against possible difficulties with infinite dimensionality. Many results in Probability and Statistics rely on Hilbert space methods: information inequalities; the Blackwell-Rao theorem; the construction of densities and abstract conditional expectations; Hellinger differentiability; prediction in time series; Gaussian process theory; martingale theory; stochastic integration; and much more.

Some of the basic theory for Hilbert space is established in Appendix B. For the next several Chapters, the following two Hilbert space results, specialized to L^2 spaces, will suffice.

(1) **Cauchy-Schwarz inequality:** $|\mu(fg)| \le \|f\|_2 \|g\|_2$ for all f, g in $L^2(\mu)$, which follows from the Hölder inequality (Problem [15]).

(2) **Orthogonal projections:** Let $\mathcal{H}_0$ be a closed subspace of $\mathcal{L}^2(\mu)$. For each f in $\mathcal{L}^2$ there is a f_0 in $\mathcal{H}_0$, the (orthogonal) projection of f onto $\mathcal{H}_0$, for which $f - f_0$ is orthogonal to $\mathcal{H}_0$, that is, $\langle f - f_0, g\rangle = 0$ for all g in $\mathcal{H}_0$. The point f_0 minimizes $\|f - h\|$ over all h in $\mathcal{H}_0$. The projection f_0 is unique up to a μ-almost sure equivalence.

REMARK. A closed subspace $\mathcal{H}_0$ of $\mathcal{L}^2$ must contain all f in $\mathcal{L}^2$ for which there exist $f_n \in \mathcal{H}_0$ with $\|f_n - f\|_2 \to 0$. In particular, if f belongs to $\mathcal{H}_0$ and $g = f$ a.e. $[\mu]$ then g must also belong to $\mathcal{H}_0$. If $\mathcal{H}_0$ is closed, the set of equivalence classes $\widetilde{\mathcal{H}}_0 = \{[f] : f \in \mathcal{H}_0\}$ must be a closed subspace of $L^2(\mu)$, and $\mathcal{H}_0$ must equal the union of all equivalence classes in $\widetilde{\mathcal{H}}_0$.

For us the most important subspaces of $\mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$ will be defined by the sub-sigma-fields $\mathcal{A}_0$ of $\mathcal{A}$. Let $\mathcal{H}_0 = \mathcal{L}^2(\mathcal{X}, \mathcal{A}_0, \mu)$. The corresponding $L^2(\mathcal{X}, \mathcal{A}_0, \mu)$ is a Hilbert space in its own right, and therefore it is a closed subspace of $L^2(\mathcal{X}, \mathcal{A}, \mu)$. Consequently $\mathcal{H}_0$ is a complete subspace of $\mathcal{L}^2$: if $\{f_n\}$ is a Cauchy sequence in $\mathcal{H}_0$ then there exists an $f_0 \in \mathcal{H}_0$ such that $\|f_n - f_0\|_2 \to 0$. However, $\{f_n\}$ also converges to every other $\mathcal{A}$-measurable f for which $f = f_0$ a.e. $[\mu]$. Unless $\mathcal{A}_0$ contains all μ-negligible sets from $\mathcal{A}$, the limit f need not be $\mathcal{A}_0$-measurable; the subspace $\mathcal{H}_0$ need not be closed. If we work instead with the corresponding $L^2(\mathcal{X}, \mathcal{A}, \mu)$ and $L^2(\mathcal{X}, \mathcal{A}_0, \mu)$ we do get a closed subspace, because the equivalence class of the limit function is uniquely determined.

*8. Uniform integrability

Suppose $\{f_n\}$ is a sequence of measurable functions converging almost surely to a limit f. If the sequence is dominated by some μ-integrable function F, then $2F \ge |f_n - f| \to 0$ almost surely, from which it follows, via Dominated Convergence, that $\mu|f_n - f| \to 0$. That is, domination plus almost sure convergence imply convergence in $\mathcal{L}^1(\mu)$ norm. The converse is not true: μ equal to Lebesgue measure and $f_n(x) := n\{(n+1)^{-1} < x \le n^{-1}\}$ provides an instance of $\mathcal{L}^1$ convergence without domination.

At least when we deal with finite measures, there is an elegant circle of equivalences, involving a concept (convergence in measure) slightly weaker than almost sure convergence and a concept (uniform integrability) slightly weaker than domination. With no loss of generality, I will explain the connections for a sequence of random variables $\{X_n\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The sequence is said to ***converge in probability*** to a random variable X, sometimes written $X_n \xrightarrow{\mathbb{P}} X$, if $\mathbb{P}\{|X_n - X| > \epsilon\} \to 0$ for each $\epsilon > 0$. Problem [14] guides you through the proofs of the following facts.

(a) If $\{X_n\}$ converges to X almost surely then $X_n \to X$ in probability, but the converse is false: there exist sequences that converge in probability but not almost surely.

(b) If $\{X_n\}$ converges in probability to X, there is an increasing sequence of positive integers $\{n(k)\}$ for which $\lim_{k\to\infty} X_{n(k)} = X$ almost surely.

If a random variable Z is integrable then a Dominated Convergence argument shows that $\mathbb{P}|Z|\{|Z| > M\} \to 0$ as $M \to \infty$. Uniform integrability requires that the convergence holds uniformly over a class of random variables. Very roughly speaking, it lets us act almost as if all the random variables were bounded by a constant M, at least as far as $\mathcal{L}^1$ arguments are concerned.

<30> **Definition.** *A family of random variables $\{Z_t : t \in T\}$ is said to be uniformly integrable if $\sup_{t\in T} \mathbb{P}|Z_t|\{|Z_t| > M\} \to 0$ as $M \to \infty$.*

It is sometimes slightly more convenient to check for uniform integrability by means of an ϵ-δ characterization.

<31> **Lemma.** *A family of random variables $\{Z_t : t \in T\}$ is uniformly integrable if and only if both the following conditions hold:*

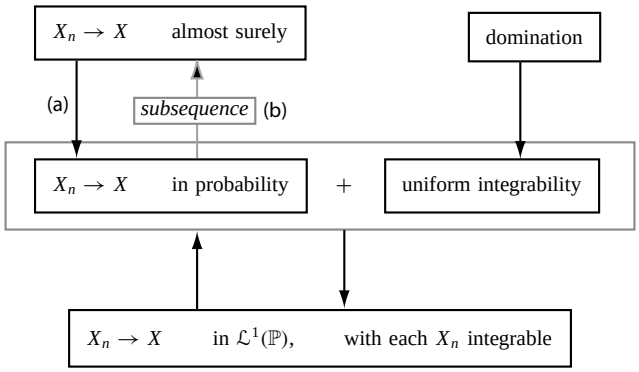
(i) $\sup_{t\in T} \mathbb{P}|Z_t| < \infty$

(ii) for each $\epsilon > 0$ there exists a $\delta > 0$ such that $\sup_{t\in T} \mathbb{P}|Z_t|F \le \epsilon$ for every event F with $\mathbb{P}F < \delta$.

REMARK. Requirement (i) is superfluous if, for each $\delta > 0$, the space Ω can be partitioned into finitely many pieces each with measure less than δ.

Proof. Given uniform integrability, (i) follows from $\mathbb{P}|Z_t| \le M + \mathbb{P}|Z_t|\{|Z_t| > M\}$, and (ii) follows from $\mathbb{P}|Z_t|F \le M\mathbb{P}F + \mathbb{P}|Z_t|\{|Z_t| > M\}$.

Conversely, if (i) and (ii) hold then the event $\{|Z_t| > M\}$ is a candidate for the F in (ii) when M is so large that $\mathbb{P}\{|Z_t| > M\} \le \sup_{t\in T} \mathbb{P}|Z_t|/M < \delta$. It follows that $\sup_{t\in T} \mathbb{P}|Z_t|\{|Z_t| > M\} \le \epsilon$ if M is large enough. □



The diagram summarizes the interconnections between the various convergence concepts, with each arrow denoting an implication. The relationship between almost sure convergence and convergence in probability corresponds to results (a) and (b) noted above. A family $\{Z_t : t \in T\}$ dominated by an integrable random variable Y

is also uniformly integrable, because $Y\{Y \geq M\} \geq |Z_t|\{|Z_t| \geq M\}$ for every t. Only the implications leading to and from the box for the $\mathcal{L}^1$ convergence remain to be proved.

<32> **Theorem.** *Let $\{X_n : n \in \mathbb{N}\}$ be a sequence of integrable random variables. The following two conditions are equivalent.*

(i) The sequence is uniformly integrable and it converges in probability to a random variable X_∞, which is necessarily integrable.

(ii) The sequence converges in $\mathcal{L}^1$ norm, $\mathbb{P}|X_n - X_\infty| \to 0$, with a limit X_∞ that is necessarily integrable.

Proof. Suppose (i) holds. The assertion about integrability of X_∞ follows from Fatou's lemma, because $|X_{n'}| \to |X_\infty|$ almost surely along some subsequence, so that $\mathbb{P}|X_\infty| \leq \liminf_{n'} \mathbb{P}|X_{n'}| \leq \sup_n \mathbb{P}|X_n| < \infty$. To prove $\mathcal{L}^1$ convergence, first split according to whether $|X_n - X_\infty|$ is less than ϵ or not, and then split according to whether $\max(|X_n|, |X_\infty|)$ is less than some large constant M or not.

$$\begin{aligned}\mathbb{P}|X_n - X_\infty| &\leq \epsilon + \mathbb{P}\left(|X_n| + |X_\infty|\right)\{|X_n - X_\infty| > \epsilon\} \\ &\leq \epsilon + 2M\mathbb{P}\{|X_n - X_\infty| > \epsilon\} + \mathbb{P}\left(|X_n| + |X_\infty|\right)\{|X_n| \vee |X_\infty| > M\}.\end{aligned}$$

Split the event $\{|X_n| \vee |X_\infty| > M\}$ according to which of the two random variables is larger, to bound the last term by $2\mathbb{P}|X_n|\{|X_n| > M\} + 2\mathbb{P}|X_\infty|\{|X_\infty| > M\}$. Invoke uniform integrability of $\{X_n\}$ and integrability of X_∞ to find an M that makes this bound small, uniformly in n. With M fixed, the convergence in probability sends $M\mathbb{P}\{|X_n - X_\infty| > \epsilon\}$ to zero as $n \to \infty$.

Conversely, if the sequence converges in $\mathcal{L}^1$, then X_∞ must be integrable, because $\mathbb{P}|X_\infty| \leq \mathbb{P}|X_n| + \mathbb{P}|X_n - X_\infty|$ for each n. When $|X_n| \leq M$ or $|X_\infty| > M/2$, the inequality

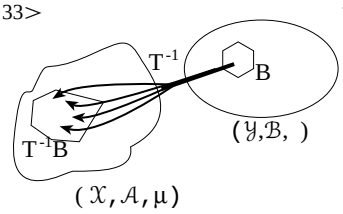
$$|X_n|\{|X_n| > M\} \leq |X_\infty|\{|X_\infty| > M/2\} + 2|X_n - X_\infty|,$$

is easy to check; and when $|X_\infty| \leq M/2$ and $|X_n| > M$, it follows from the inequality $|X_n - X_\infty| \geq |X_n| - |X_\infty| \geq |X_n|/2$. Take expectations, choose M large enough to make the contribution from X_∞ small, then let n tend to infinity to find an n_0 such that $\mathbb{P}|X_n|\{|X_n| > M\} < \epsilon$ for $n > n_0$. Increase M if necessary to handle the corresponding tail contributions for $n \leq n_0$. □

9. Image measures and distributions

Suppose μ is a measure on a sigma-field $\mathcal{A}$ of subsets of $\mathcal{X}$ and T is a map from $\mathcal{X}$ into a set $\mathcal{Y}$, equipped with a sigma-field $\mathcal{B}$. If T is $\mathcal{A}\backslash\mathcal{B}$-measurable we can carry μ over to $\mathcal{Y}$, by defining

<33> $$\nu B := \mu(T^{-1}B) \qquad \text{for each } B \text{ in } \mathcal{B}.$$



Actually the operation is more one of carrying the sets back to the measure rather than carrying the measure over to the sets, but the net result is the definition of a new set function on $\mathcal{B}$.

It is easy to check that ν is a measure on $\mathcal{B}$, using facts such as $T^{-1}(B^c) = \left(T^{-1}B\right)^c$ and $T^{-1}(\cup_i B_i) = \cup_i T^{-1}B_i$. It is called the ***image measure of*** μ ***under*** T, or just the image measure, and is denoted by μT^{-1} or μ_T or $T(\mu)$, or even just $T\mu$. The third and fourth forms, which I prefer to use, have the nice property that if μ is a point mass concentrated at x then $T(\mu)$ denotes a point mass concentrated at $T(x)$.

Starting from definition <33> we could prove facts about integrals with respect to the image measure ν. For example, we could show

<34> $$\nu g = \mu(g \circ T) \qquad \text{for all } g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B}).$$

The small circle symbol $\circ$ denotes the composition of functions: $(g \circ T)(x) := g(Tx)$.

The proof of <34> could follow the traditional path: first argue by linearity from <33> to establish the result for simple functions; then take monotone limits of simple functions to extend to $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$.

There is another method for constructing image measures that gets <34> all in one step. *Define* an increasing linear functional ν on $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ by $\nu g := \mu(g \circ T)$. It inherits the Monotone Convergence property directly from μ, because, if $0 \le g_n \uparrow g$ then $0 \le g_n \circ T \uparrow g \circ T$. By Theorem <13> it corresponds to a uniquely determined measure on $\mathcal{B}$. When restricted to indicator functions of measurable sets the new measure coincides with the measure defined by <33>, because if g is the indicator function of B then $g \circ T$ is the indicator function of $T^{-1}B$. (Why?) We have gained a theorem with almost no extra work, by starting with the linear functional as the definition of the image measure.

Using the notation $T\mu$ for image measure, we could rewrite the defining equality as $(T\mu)(g) := \mu(g \circ T)$ at least for all $g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$, a relationship that I find easier to remember.

> REMARK. In the last sentence I used the qualifier *at least*, as a reminder that the equality could easily be extended to other cases. For example, by splitting into positive and negative parts then subtracting, we could extend the equality to functions in $\mathcal{L}^1(\mathcal{Y}, \mathcal{B}, \nu)$. And so on.

Several familiar probabilistic objects are just image measures. If X is a random variable, the image measure $X(\mathbb{P})$ on $\mathcal{B}(\mathbb{R})$ is often written $\mathbb{P}_X$, and is called the ***distribution of*** **X**. More generally, if X and Y are random variables defined on the same probability space, they together define a ***random vector***, a (measurable—see Chapter 4) map $T(\omega) = (X(\omega), Y(\omega))$ from Ω into $\mathbb{R}^2$. The image measure $T(\mathbb{P})$ on $\mathcal{B}(\mathbb{R}^2)$ is called the ***joint distribution of*** **X** ***and*** **Y**, and is often denoted by $\mathbb{P}_{X,Y}$. Similar terminology applies for larger collections of random variables.

Image measures also figure in a construction that is discussed nonrigorously in many introductory textbooks. Let P be a probability measure on $\mathcal{B}(\mathbb{R})$. Its ***distribution function*** (also known as a cumulative distribution function) is defined by $F_P(x) := P(-\infty, x]$ for $x \in \mathbb{R}$. Don't confuse *distribution*, as a synonym for probability measure, with *distribution function*, which is a function derived from the measures of a particular collection of sets. The distribution function has the following properties.

(a) It is increasing, with $\lim_{x \to -\infty} F_P(x) = 0$ and $\lim_{x \to \infty} F_P(x) = 1$.

(b) It is continuous from the right: to each $\epsilon > 0$ and $x \in \mathbb{R}$, there exists a $\delta > 0$ such that $F_P(x) \le F_P(y) \le F_P(x) + \epsilon$ for $x \le y \le x + \delta$.

Property (a) follows from that fact that the integral is an increasing functional, and from Dominated Convergence applied to the sequences $(-\infty, -n] \downarrow \emptyset$ and $(-\infty, n] \uparrow \mathbb{R}$ as $n \to \infty$. Property (b) also follows from Dominated Convergence, applied to the sequence $(-\infty, x + 1/n] \downarrow (-\infty, x]$.

Except in introductory textbooks, and in works dealing with the order properties of the real line (such as the study of ranks and order statistics), distribution functions have a reduced role to play in modern probability theory, mostly in connection with the following method for building measures on $\mathcal{B}(\mathbb{R})$ as images of Lebesgue measure. In probability theory the construction often goes by the name of ***quantile transformation***.

<35> **Example.** There is a converse to the assertions (a) and (b) about distribution functions. Suppose F is a right-continuous, increasing function on $\mathbb{R}$ for which $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Then there exists a probability measure P such that $P(-\infty, x] = F(x)$ for all real x. To construct such a P, consider the ***quantile function*** q, defined by $q(t) := \inf\{x : F(x) \ge t\}$ for $0 < t < 1$.

By right continuity of the increasing function F, the set $\{x \in \mathbb{R} : F(x) \ge t\}$ is a closed interval of the form $[\alpha, \infty)$, with $\alpha = q(t)$. That is, for all $x \in \mathbb{R}$ and all $t \in (0, 1)$,

<36> $$F(x) \ge t \qquad \text{if and only if} \qquad x \ge q(t).$$

In general there are many plausible, but false, equalities related to <36>. For example, it is not true in general that $F(q(t)) = t$. However, if F is continuous and strictly increasing, then q is just the inverse function of F, and the plausible equalities hold.

Let $\mathfrak{m}$ denote Lebesgue measure restricted to the Borel sigma-field on $(0, 1)$. The image measure $P := q(\mathfrak{m})$ has the desired property,

$$P(-\infty, x] = \mathfrak{m}\{t : q(t) \le x\} = \mathfrak{m}\{t : t \le F(x)\} = F(x),$$

the first equality by definition of the image measure, and the second by equality <36>. The result is often restated as: *if ξ has a Uniform$(0, 1)$ distribution then $q(\xi)$ has distribution function F.* □

10. Generating classes of sets

To prove that all sets in a sigma-field $\mathcal{A}$ have some property one often resorts to a generating-class argument. The simplest form of such an argument has three steps:

(i) Show that all members of a subclass $\mathcal{E}$ have the property.

(ii) Show that $\mathcal{A} \subseteq \sigma(\mathcal{E})$.

(iii) Show that $\mathcal{A}_0 := \{A \in \mathcal{A} : A \text{ has the property }\}$ is a sigma-field.

Then one deduces that $\mathcal{A}_0 = \sigma(\mathcal{A}_0) \supseteq \sigma(\mathcal{E}) \supseteq \mathcal{A}$, whence $\mathcal{A}_0 = \mathcal{A}$. That is, the property holds for all sets in $\mathcal{A}$.

For some properties, direct verification of all the sigma-field requirements for $\mathcal{A}_0$ proves too difficult. In such situations an indirect argument sometimes succeeds if $\mathcal{E}$ has some extra structure. For example, if is possible to establish that $\mathcal{A}_0$ is a ***λ-system of sets***, then one needs only check one extra requirement for $\mathcal{E}$ in order to produce a successful generating-class argument.

<37> **Definition.** *A class $\mathcal{D}$ of subsets of $\mathcal{X}$ is called a λ-system if*

(i) $\mathcal{X} \in \mathcal{D}$,

(ii) if $D_1, D_2 \in \mathcal{D}$ *and* $D_1 \supseteq D_2$ *then* $D_1 \backslash D_2 \in \mathcal{D}$,

(iii) if $\{D_n\}$ *is an increasing sequence of sets in* $\mathcal{D}$ *then* $\cup_1^\infty D_n \in \mathcal{D}$.

REMARK. Some authors start from a slightly different definition, replacing requirement (iii) by

(iii)′ if $\{D_n\}$ is a sequence of disjoint sets in $\mathcal{D}$ then $\cup_1^\infty D_n \in \mathcal{D}$.

The change in definition would have little effect on the role played by λ-systems.

Many authors (including me, until recently) use the name ***Dynkin class*** instead of λ-system, but the name ***Sierpiński class*** would be more appropriate. See the Notes at the end of this Chapter.

Notice that a λ-system is also a sigma-field if and only if it is stable under finite intersections. This stability property can be inherited from a subclass $\mathcal{E}$, as in the next Theorem, which is sometimes referred to as the π–λ theorem. The π stands for *product*, an indirect reference to the stability of the subclass $\mathcal{E}$ under finite intersections (products). I think that the letter λ stands for *limit*, an indirect reference to property (iii).

<38> **Theorem.** *If $\mathcal{E}$ is stable under finite intersections, and if $\mathcal{D}$ is a λ-system with $\mathcal{D} \supseteq \mathcal{E}$, then $\mathcal{D} \supseteq \sigma(\mathcal{E})$.*

Proof. It would be enough to show that $\mathcal{D}$ is a sigma-field, by establishing that it is stable under finite intersections, but that is a little more than I know how to do. Instead we need to work with a (possibly) smaller λ-system $\mathcal{D}_0$, with $\mathcal{D} \supseteq \mathcal{D}_0 \supseteq \mathcal{E}$, for which generating class arguments can extend the assumption

<39> $$E_1 E_2 \in \mathcal{E} \qquad \text{for all } E_1, E_2 \text{ in } \mathcal{E}$$

to an assertion that

<40> $$D_1 D_2 \in \mathcal{D}_0 \qquad \text{for all } D_1, D_2 \text{ in } \mathcal{D}_0.$$

It will then follow that $\mathcal{D}_0$ is a sigma-field, which contains $\mathcal{E}$, and hence $\mathcal{D}_0 \supseteq \sigma(\mathcal{E})$.

The choice of $\mathcal{D}_0$ is easy. Let $\{\mathcal{D}_\alpha : \alpha \in A\}$ be the collection of all λ-systems with $\mathcal{D}_\alpha \supseteq \mathcal{E}$, one of them being the $\mathcal{D}$ we started with. Let $\mathcal{D}_0$ equal the intersection of all these $\mathcal{D}_\alpha$. That is, let $\mathcal{D}_0$ consist of all sets D for which $D \in \mathcal{D}_\alpha$ for each α. I leave it to you to check the easy details that prove $\mathcal{D}_0$ to be a λ-system. In other words, $\mathcal{D}_0$ is the smallest λ-system containg $\mathcal{E}$; it is the λ-system generated by $\mathcal{E}$.

To upgrade <39> to <40> we have to replace each E_i on the left-hand side by a D_i in $\mathcal{D}_0$, without going outside the class $\mathcal{D}_0$. The trick is to work one component at a time. Start with the E_1. Define $\mathcal{D}_1 := \{A : AE \in \mathcal{D}_0 \text{ for each } E \in \mathcal{E}\}$. From <39>, we have $\mathcal{D}_1 \supseteq \mathcal{E}$. If we show that $\mathcal{D}_1$ is a λ-system then it will follow that $\mathcal{D}_1 \supseteq \mathcal{D}_0$, because $\mathcal{D}_0$ is the smallest λ-system containing $\mathcal{E}$. Actually, the assertion that $\mathcal{D}_1$ is

λ-system is trivial; it follows immediately from the λ-system properties for $\mathcal{D}_0$ and identities like $(A_1 \backslash A_2)E = (A_1E)\backslash(A_2E)$ and $(\cup_i A_i)\,E = \cup(A_iE)$.

The inclusion $\mathcal{D}_1 \supseteq \mathcal{D}_0$ implies that $D_1E_2 \in \mathcal{D}_0$ for all $D_1 \in \mathcal{D}_0$ and all $E_2 \in \mathcal{E}$. Put another way—this step is the only subtlety in the proof—we can assert that the class $\mathcal{D}_2 := \{B : BD \in \mathcal{D}_0 \text{ for each } D \in \mathcal{D}_0\}$ contains $\mathcal{E}$. Just write D_1 instead of D, and E_1 instead of B, in the definition to see that it is only a matter of switching the order of the sets.

Argue in the same way as for $\mathcal{D}_1$ to show that $\mathcal{D}_2$ is also a λ-system. It then follows that $\mathcal{D}_2 \supseteq \mathcal{D}_0$, which is another way of expressing assertion <40>. □

The proof of the last Theorem is typical of many generating class arguments, in that it is trivial once one knows what one has to check. The Theorem, or its analog for classes of functions (see the next Section), will be my main method for establishing sigma-field properties. You will be getting plenty of practice at filling in the details behind frequent assertions of "a generating class argument shows that ...." Here is a typical example to get you started.

<41> **Exercise.** Let μ and ν be finite measures on $\mathcal{B}(\mathbb{R})$ with the same distribution function. That is, $\mu(-\infty, t] = \nu(-\infty, t]$ for all real t. Show that $\mu B = \nu B$ for all $B \in \mathcal{B}(\mathbb{R})$, that is, $\mu = \nu$ as Borel measures.

SOLUTION: Write $\mathcal{E}$ for the class of all intervals $(-\infty, t]$, with $t \in \mathbb{R}$. Clearly $\mathcal{E}$ is stable under finite intersections. From Example <4>, we know that $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$. It is easy to check that the class $\mathcal{D} := \{B \in \mathcal{B}(\mathbb{R}) : \mu B = \nu B\}$ is a λ-system. For example, if $B_n \in \mathcal{D}$ and $B_n \uparrow B$ then $\mu B = \lim_n \mu B_n = \lim_n \nu B_n = \nu B$, by Monotone Convergence. It follows from Theorem <38> that $\mathcal{D} \supseteq \sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$, and the equality of the two Borel measures is established. □

When you employ a λ-system argument be sure to verify the properties required of $\mathcal{E}$. The next Example shows what can happen if you forget about the stability under finite intersections.

<42> **Example.** Consider a set $\mathcal{X}$ consisting of four points, labelled nw, ne, sw, and se. Let $\mathcal{E}$ consist of $\mathcal{X}$ and the subsets N = {nw, ne}, S = {sw, se}, E = {ne, se}, and W = {nw, sw}. Notice that $\mathcal{E}$ generates the sigma-field of all subsets of $\mathcal{X}$, but it is not stable under finite intersections. Let μ and ν be probability measures for which

$\mu(\text{nw}) = 1/2$	$\mu(\text{ne}) = 0$	$\nu(\text{nw}) = 0$	$\nu(\text{ne}) = 1/2$
$\mu(\text{sw}) = 0$	$\mu(\text{se}) = 1/2$	$\nu(\text{sw}) = 1/2$	$\nu(\text{se}) = 0$

Both measures give the the value $1/2$ to each of N, S, E, and W, but they differ in the values they give to the four singletons. □

*11. Generating classes of functions

Theorem <38> is often used as the starting point for proving facts about measurable functions. One first invokes the Theorem to establish a property for sets in a sigma-field, then one extends by taking limits of simple functions to $\mathcal{M}^+$ and beyond, using Monotone Convergence and linearity arguments. Sometimes it is simpler to invoke an analog of the λ-system property for classes of functions.

<43> **Definition.** *Call a class $\mathcal{H}^+$ of bounded, nonnegative functions on a set $\mathcal{X}$ a* ***λ-cone*** *if:*

(i) $\mathcal{H}^+$ is a cone, that is, if $h_1, h_2 \in \mathcal{H}^+$ and α_1 and α_2 are nonnegative constants then $\alpha_1 h_1 + \alpha_2 h_2 \in \mathcal{H}^+$;

(ii) each nonnegative constant function belongs to $\mathcal{H}^+$;

(iii) if $h_1, h_2 \in \mathcal{H}^+$ and $h_1 \geq h_2$ then $h_1 - h_2 \in \mathcal{H}^+$;

(iv) if $\{h_n\}$ is an increasing sequence of functions in $\mathcal{H}^+$ whose pointwise limit h is bounded then $h \in \mathcal{H}^+$.

Typically $\mathcal{H}^+$ consists of the nonnegative functions in a vector space of bounded functions that is stable under pairwise maxima and minima.

REMARK. The name λ-cone is not standard. I found it hard to come up with a name that was both suggestive of the defining properties and analogous to the name for the corresponding classes of sets. For a while I used the term Dynkin-cones but abandoned it for historical reasons. (See the Notes.) I also toyed with the name cdl-cone, as a reminder that the cone contains the (positive) constant functions and that it is stable under (proper) differences and (monotone increasing) limits of uniformly bounded sequences.

The sigma-field properties of λ-cones are slightly harder to establish than their λ-system analogs, but the reward of more streamlined proofs will make the extra, one-time effort worthwhile. First we need an analog of the fact that a λ-system that is stable under finite intersections is also a sigma-field.

<44> **Lemma.** *If a λ-cone $\mathcal{H}^+$ is stable under the formation of pointwise products of pairs of functions then it consists of all bounded, nonnegative, $\sigma(\mathcal{H}^+)$-measurable functions, where $\sigma(\mathcal{H}^+)$ denotes the sigma-field generated by $\mathcal{H}^+$.*

Proof. First note that $\mathcal{H}^+$ must be stable under uniform limits. For suppose $h_n \to h$ uniformly, with $h_n \in \mathcal{H}^+$. Write δ_n for 2^{-n}. With no loss of generality we may suppose $h_n + \delta_n \geq h \geq h_n - \delta_n$ for all n. Notice that

$$h_n + 3\delta_n = h_n + \delta_n + \delta_{n-1} \geq h + \delta_{n-1} \geq h_{n-1}.$$

From the monotone convergence, $0 \leq h_n + 3(\delta_1 + \ldots + \delta_n) \uparrow h + 3$, deduce that $h + 3 \in \mathcal{H}^+$, and hence, via the proper difference property (iii), $h \in \mathcal{H}^+$.

Via uniform limits we can now show that $\mathcal{H}^+$ is stable under composition with any continuous nonnegative function f. Let h be a member of $\mathcal{H}^+$, bounded above by a constant D. By a trivial generalization of Problem [25], there exists a sequence of polynomials $p_n(\cdot)$ such that $\sup_{0 \leq t \leq D} |p_n(t) - f(t)| < 1/n$. The function $f_n(h) := p_n(h) + 1/n$ takes only nonegative values, and it converges uniformly to $f(h)$. Suppose $f_n(t) = a_0 + a_1 t + \ldots + a_k t^k$. Then

$$f_n(h) = (a_0^+ + a_1^+ h + \ldots + a_k^+ h^k) - (a_0^- + a_1^- h + \ldots + a_k^- h^k) \geq 0.$$

By virtue of properties (i) and (ii) of λ-cones, and the assumed stability under products, both terms on the right-hand side belong to $\mathcal{H}^+$. The proper differencing property then gives $f_n(h) \in \mathcal{H}^+$. Pass uniformly to the limit to get $f(h) \in \mathcal{H}^+$.

Write $\mathcal{E}$ for the class of all sets of the form $\{h < C\}$, with $h \in \mathcal{H}^+$ and C a positive constant. From Example <7>, every h in $\mathcal{H}^+$ is $\sigma(\mathcal{E})$-measurable, and

hence $\sigma(\mathcal{E}) = \sigma(\mathcal{H}^+)$. For a fixed h and C, the continuous function $(1-(h/C)^n)^+$ of h belongs to $\mathcal{H}^+$, and it increases monotonely to the indicator of $\{h < C\}$. Thus the indicators of all sets in $\mathcal{E}$ belong to $\mathcal{H}^+$. The assumptions about $\mathcal{H}^+$ ensure that the class $\mathcal{B}$ of all sets whose indicator functions belong to $\mathcal{H}^+$ is stable under finite intersections (products), complements (subtract from 1), and increasing countable unions (montone increasing limits). That is, $\mathcal{B}$ is a λ-system, stable under finite intersections, and containing $\mathcal{E}$. It is a sigma-field containing $\mathcal{E}$. Thus $\mathcal{B} \supseteq \sigma(\mathcal{E}) = \sigma(\mathcal{H}^+)$. That is, $\mathcal{H}^+$ contains all indicators of sets in $\sigma(\mathcal{H}^+)$.

Finally, let k be a bounded, nonnegative, $\sigma(\mathcal{H}^+)$-measurable function. From the fact that each of the sets $\{k \geq i/2^n\}$, for $i = 1, \ldots, 4^n$, belongs to the cone $\mathcal{H}^+$, we have $k_n := 2^{-n}\sum_{i=1}^{4^n}\{k \geq i/2^n\} \in \mathcal{H}^+$. The functions k_n increase monotonely to k, which consequently also belongs to $\mathcal{H}^+$. □

<45> **Theorem.** *Let $\mathcal{H}^+$ be a λ-cone of bounded, nonnegative functions, and $\mathcal{G}$ be a subclass of $\mathcal{H}^+$ that is stable under the formation of pointwise products of pairs of functions. Then $\mathcal{H}^+$ contains all bounded, nonnegative, $\sigma(\mathcal{G})$-measurable functions.*

Proof. Let $\mathcal{H}_0^+$ be the smallest λ-cone containing $\mathcal{G}$. From the previous Lemma, it is enough to show that $\mathcal{H}_0^+$ is stable under pairwise products.

Argue as in Theorem <38> for λ-systems of sets. A routine calculation shows that $\mathcal{H}_1^+ := \{h \in \mathcal{H}_0^+ : hg \in \mathcal{H}_0^+ \text{ for all } g \text{ in } \mathcal{G}\}$ is a λ-cone containing $\mathcal{G}$, and hence $\mathcal{H}_1^+ = \mathcal{H}_0^+$. That is, $h_0 g \in \mathcal{H}_0^+$ for all $h_0 \in \mathcal{H}_0^+$ and $g \in \mathcal{G}$. Similarly, the class $\mathcal{H}_2^+ := \{h \in \mathcal{H}_0^+ : h_0 h \in \mathcal{H}_0^+ \text{ for all } h_0 \text{ in } \mathcal{H}_0^+\}$ is a λ-cone. By the result for $\mathcal{H}_1^+$ we have $\mathcal{H}_2^+ \supseteq \mathcal{G}$, and hence $\mathcal{H}_2^+ = \mathcal{H}_0^+$. That is, $\mathcal{H}_0^+$ is stable under products. □

<46> **Exercise.** Let μ be a finite measure on $\mathcal{B}(\mathbb{R}^k)$. Write $\mathbb{C}_0$ for the vector space of all continuous real functions on $\mathbb{R}^k$ with compact support. Suppose f belongs to $\mathcal{L}^1(\mu)$. Show that for each $\epsilon > 0$ there exists a g in $\mathbb{C}_0$ such that $\mu|f - g| < \epsilon$. That is, show that $\mathbb{C}_0$ is dense in $\mathcal{L}^1(\mu)$ under its $\mathcal{L}^1$ norm. □

SOLUTION: Define $\mathcal{H}$ as the collection of all bounded functions in $\mathcal{L}^1(\mu)$ that can be approximated arbitrarily closely by functions from $\mathbb{C}_0$. Check that the class $\mathcal{H}^+$ of nonnegative functions in $\mathcal{H}$ is a λ-cone. Trivially it contains $\mathbb{C}_0^+$, the class of nonnegative members of $\mathbb{C}_0$. The sigma-field $\sigma(\mathbb{C}_0^+)$ coincides with the Borel sigma-field. Why? The class $\mathcal{H}^+$ consists of all bounded, nonnegative Borel measurable functions.

To approximate a general f in $\mathcal{L}^1(\mu)$, first reduce to the case of nonnegative functions by splitting into positive and negative parts. Then invoke Dominated Convergence to find a finite n for which $\mu|f^+ - f^+ \wedge n| < \epsilon$, then approximate $f^+ \wedge n$ by a member of $\mathbb{C}_0^+$. See Problem [26] for the extension of the approximation result to infinite measures. □

12. Problems

[1] Suppose events $A_1, A_2, \ldots$, in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, are independent: meaning that $\mathbb{P}(A_{i_1}A_{i_2}\ldots A_{i_k}) = \mathbb{P}A_{i_1}\mathbb{P}A_{i_2}\ldots\mathbb{P}A_{i_k}$ for all choices of distinct subscripts $i_1, i_2, \ldots, i_k$, all k. Suppose $\sum_{i=1}^{\infty}\mathbb{P}A_i = \infty$.

(i) Using the inequality $e^{-x} \geq 1 - x$, show that

$$\mathbb{P} \max_{n \leq i \leq m} A_i = 1 - \prod_{n \leq i \leq m} (1 - \mathbb{P}A_i) \geq 1 - \exp\left(- \sum_{n \leq i \leq m} \mathbb{P}A_i \right)$$

(ii) Let m then n tend to infinity, to deduce (via Dominated Convergence) that $\mathbb{P} \limsup_i A_i = 1$. That is, $\mathbb{P}\{A_i \text{ i. o.}\} = 1$.

REMARK. The result gives a converse for the Borel-Cantelli lemma from Example <29>. The next Problem establishes a similar result under weaker assumptions.

[2] Let $A_1, A_2, \ldots$ be events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define $X_n = A_1 + \ldots + A_n$ and $\sigma_n = \mathbb{P}X_n$. Suppose $\sigma_n \to \infty$ and $\|X_n/\sigma_n\|_2 \to 1$. *(Compare with the inequality* $\|X_n/\sigma_n\|_2 \geq 1$, *which follows from Jensen's inequality.)*

(i) Show that

$$\{X_n = 0\} \leq \frac{(k - X_n)(k + 1 - X_n)}{k(k+1)}$$

for each positive integer k.

(ii) By an appropriate choice of k (depending on n) in (i), deduce that $\sum_1^\infty A_i \geq 1$ almost surely.

(iii) Prove that $\sum_m^\infty A_i \geq 1$ almost surely, for each fixed m. Hint: Show that the two convergence assumptions also hold for the sequence $A_m, A_{m+1}, \ldots$.

(iv) Deduce that $\mathbb{P}\{\omega \in A_i \text{ i. o. }\} = 1$.

(v) If $\{B_i\}$ is a sequence of events for which $\sum_i \mathbb{P}B_i = \infty$ and $\mathbb{P}B_iB_j = \mathbb{P}B_i\mathbb{P}B_j$ for $i \neq j$, show that $\mathbb{P}\{\omega \in B_i \text{ i. o. }\} = 1$.

[3] Suppose T is a function from a set $\mathcal{X}$ into a set $\mathcal{Y}$, and suppose that $\mathcal{Y}$ is equipped with a σ-field $\mathcal{B}$. Define $\mathcal{A}$ as the sigma-field of sets of the form $T^{-1}B$, with B in $\mathcal{B}$. Suppose $f \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$. Show that there exists a $\mathcal{B}\backslash\mathcal{B}[0,\infty]$-measurable function g from $\mathcal{Y}$ into $[0,\infty]$ such that $f(x) = g(T(x))$, for all x in $\mathcal{X}$, by following these steps.

(i) Show that $\mathcal{A}$ is a σ-field on $\mathcal{X}$. (It is called the σ-field generated by the map T. It is often denoted by $\sigma(T)$.)

(ii) Show that $\{f \geq i/2^n\} = T^{-1}B_{i,n}$ for some $B_{i,n}$ in $\mathcal{B}$. Define

$$f_n = 2^{-n} \sum_{i=1}^{4^n} \{f \geq i/2^n\} \qquad \text{and} \qquad g_n = 2^{-n} \sum_{i=1}^{4^n} B_{i,n}.$$

Show that $f_n(x) = g_n(T(x))$ for all x.

(iii) Define $g(y) = \limsup g_n(y)$ for each y in $\mathcal{Y}$. Show that g has the desired property. (Question: Why can't we define $g(y) = \lim g_n(y)$?)

[4] Let $g_1, g_2, \ldots$ be $\mathcal{A}\backslash\mathcal{B}(\mathbb{R})$-measurable functions from $\mathcal{X}$ into $\mathbb{R}$. Show that $\{\limsup_n g_n > t\} = \bigcup_{\substack{r \in \mathbb{Q} \\ r > t}} \bigcap_{m=1}^\infty \bigcup_{i \geq m} \{g_i > r\}$. Deduce, without any appeal to Example <8>, that $\limsup g_n$ is $\mathcal{A}\backslash\mathcal{B}(\overline{\mathbb{R}})$-measurable. Warning: Be careful about

strict inequalities that turn into nonstrict inequalities in the limit—it is possible to have $x_n > x$ for all n and still have $\limsup_n x_n = x$.

[5] Suppose a class of sets $\mathcal{E}$ cannot separate a particular pair of points x, y: for every E in $\mathcal{E}$, either $\{x, y\} \subseteq E$ or $\{x, y\} \subseteq E^c$. Show that $\sigma(\mathcal{E})$ also cannot separate the pair.

[6] A collection of sets $\mathcal{F}_0$ that is stable under finite unions, finite intersections, and complements is called a field. A nonnegative set function μ defined on $\mathcal{F}_0$ is called a finitely additive measure if $\mu\left(\cup_{i\le n} F_i\right) = \sum_{i\le n} \mu F_i$ for every finite collection of disjoint sets in $\mathcal{F}_0$. The set function is said to be countably additive on $\mathcal{F}_0$ if $\mu\left(\cup_{i\in\mathbb{N}} F_i\right) = \sum_{i\in\mathbb{N}} \mu F_i$ for every countable collection of disjoint sets in $\mathcal{F}_0$ *whose union belongs to* $\mathcal{F}$. Suppose $\mu\mathcal{X} < \infty$. Show that μ is countably additive on $\mathcal{F}_0$ if and only if $\mu A_n \downarrow 0$ for every decreasing sequence in $\mathcal{F}_0$ with empty intersection. Hint: For the argument in one direction, consider the union of differences $A_i \setminus A_{i+1}$.

[7] Let $f_1, \ldots, f_n$ be functions in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$, and let μ be a measure on $\mathcal{A}$. Show that $\mu\left(\vee_i f_i\right) \le \sum_i \mu f_i \le \mu\left(\vee_i f_i\right) + \sum_{i<j} \mu\left(f_i \wedge f_j\right)$ where $\vee$ denotes pointwise maxima of functions and $\wedge$ denotes pointwise minima.

[8] Let μ be a finite measure and f be a measurable function. For each positive integer k, show that $\mu|f|^k < \infty$ if and only if $\sum_{n=1}^{\infty} n^{k-1}\mu\{|f| \ge n\} < \infty$.

[9] Suppose $\nu := T\mu$, the image of the measure μ under the measurable map T. Show that $f \in \mathcal{L}^1(\nu)$ if and only if $f \circ T \in \mathcal{L}^1(\mu)$, in which case $\nu f = \mu(f \circ T)$.

[10] Let $\{h_n\}$, $\{f_n\}$, and $\{g_n\}$ be sequences of μ-integrable functions that converge μ almost everywhere to limits h, f and g. Suppose $h_n(x) \le f_n(x) \le g_n(x)$ for all x. Suppose also that $\mu h_n \to \mu h$ and $\mu g_n \to \mu g$. Adapt the proof of Dominated Convergence to prove that $\mu f_n \to \mu f$.

[11] A collection of sets is called a monotone class if it is stable under unions of increasing sequences and intersections of decreasing sequences. Adapt the argument from Theorem <38> to prove: if a class $\mathcal{E}$ is stable under finite unions and complements then $\sigma(\mathcal{E})$ equals the smallest monotone class containing $\mathcal{E}$.

[12] Let μ be a finite measure on the Borel sigma-field $\mathcal{B}(\mathcal{X})$ of a metric space $\mathcal{X}$. Call a set B ***inner regular*** if $\mu B = \sup\{\mu F : B \supseteq F \text{ closed }\}$ and ***outer regular*** if $\mu B = \inf\{\mu F : B \subseteq G \text{ open }\}$

(i) Prove that the class $\mathcal{B}_0$ of all Borel sets that are both inner and outer regular is a sigma-field. Deduce that every Borel set is inner regular.

(ii) Suppose μ is tight: for each $\epsilon > 0$ there exists a compact K_ϵ such that $\mu K_\epsilon^c < \epsilon$. Show that the F in the definition of inner regularity can then be assumed compact.

(iii) When μ is tight, show that there exists a sequence of disjoint compacts subsets $\{K_i : i \in \mathbb{N}\}$ of $\mathcal{X}$ such that $\mu\left(\cup_i K_i\right)^c = 0$.

[13] Let μ be a finite measure on the Borel sigma-field of a complete, separable metric space $\mathcal{X}$. Show that μ is tight: for each $\epsilon > 0$ there exists a compact K_ϵ such that $\mu K_\epsilon^c < \epsilon$. Hint: For each positive integer n, show that the space $\mathcal{X}$ is a countable

union of closed balls with radius $1/n$. Find a finite family of such balls whose union B_n has μ measure greater than $\mu\mathcal{X} - \epsilon/2^n$. Show that $\cap_n B_n$ is compact, using the total-boundedness characterization of compact subsets of complete metric spaces.

[14] A sequence of random variables $\{X_n\}$ is said to ***converge in probability*** to a random variable X, written $X_n \xrightarrow{\mathbb{P}} X$, if $\mathbb{P}\{|X_n - X| > \epsilon\} \to 0$ for each $\epsilon > 0$.

(i) If $X_n \to X$ almost surely, show that $1 \geq \{|X_n - X| > \epsilon\} \to 0$ almost surely. Deduce via Dominated Convergence that X_n converges in probability to X.

(ii) Give an example of a sequence $\{X_n\}$ that converges to X in probability but not almost surely.

(iii) Suppose $X_n \to X$ in probability. Show that there is an increasing sequence of positive integers $\{n(k)\}$ for which $\sum_k \mathbb{P}\{|X_{n(k)} - X| > 1/k\} < \infty$. Deduce that $X_{n(k)} \to X$ almost surely.

[15] Let f and g be measurable functions on $(\mathcal{X}, \mathcal{A}, \mu)$, and r and s be positive real numbers for which $r^{-1} + s^{-1} = 1$. Show that $\mu|fg| \leq (\mu|f|^r)^{1/r} (\mu|g|^s)^{1/s}$ by arguing as follows. First dispose of the trivial case where one of the factors on the righthand side is 0 or ∞. Then, without loss of generality (why?), assume that $\mu|f|^r = 1 = \mu|g|^s$. Use concavity of the logarithm function to show that $|fg| \leq |f|^r/r + |g|^s/s$, and then integrate with respect to μ. *This result is called the* ***Hölder inequality***.

[16] Generalize the Hölder inequality (Problem [15]) to more than two measurable functions $f_1, \ldots, f_k$, and positive real numbers $r_1, \ldots, r_k$ for which $\sum_i r_i^{-1} = 1$. Show that $\mu|f_1 \ldots f_k| \leq \prod_i (\mu|f_i|^{r_i})^{1/r_i}$.

[17] Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, f and g be measurable functions, and r be a real number with $r \geq 1$. Define $\|f\|_r = (\mu|f|^r)^{1/r}$. Follow these steps to prove ***Minkowski's inequality***: $\|f + g\|_r \leq \|f\|_r + \|g\|_r$.

(i) From the inequality $|x + y|^r \leq |2x|^r + |2y|^r$ deduce that $\|f + g\|_r < \infty$ if $\|f\|_r < \infty$ and $\|g\|_r < \infty$.

(ii) Dispose of trivial cases, such as $\|f\|_r = 0$ or $\|f\|_r = \infty$.

(iii) For arbitrary positive constants c and d argue by convexity that

$$\left(\frac{|f| + |g|}{c + d}\right)^r \leq \frac{c}{c + d}\left(\frac{|f|}{c}\right)^r + \frac{d}{c + d}\left(\frac{|g|}{d}\right)^r$$

(iv) Integrate, then choose $c = \|f\|_r$ and $d = \|g\|_r$ to complete the proof.

[18] For f in $\mathcal{L}^1(\mu)$ define $\|f\|_1 = \mu|f|$. Let $\{f_n\}$ be a Cauchy sequence in $\mathcal{L}^1(\mu)$, that is, $\|f_n - f_m\|_1 \to 0$ as $\min(m, n) \to \infty$. Show that there exists an f in $\mathcal{L}^1(\mu)$ for which $\|f_n - f\|_1 \to 0$, by following these steps.

(i) Find an increasing sequence $\{n(k)\}$ such that $\sum_{k=1}^{\infty} \|f_{n(k)} - f_{n(k+1)}\|_1 < \infty$. Deduce that the function $H := \sum_{k=1}^{\infty} |f_{n(k)} - f_{n(k+1)}|$ is integrable.

(ii) Show that there exists a real-valued, measurable function f for which

$$H \geq |f_{n(k)}(x) - f(x)| \to 0 \qquad \text{as } k \to \infty, \text{ for } \mu \text{ almost all } x.$$

Deduce that $\|f_{n(k)} - f\|_1 \to 0$ as $k \to \infty$.

(iii) Show that f belongs to $\mathcal{L}^1(\mu)$ and $\|f_n - f\|_1 \to 0$ as $n \to \infty$.

[19] Let $\{f_n\}$ be a Cauchy sequence in $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$, that is, $\|f_n - f_m\|_p \to 0$ as $\min(m, n) \to \infty$. Show that there exists a function f in $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$ for which $\|f_n - f\|_p \to 0$, by following these steps.

(i) Find an increasing sequence $\{n(k)\}$ such that $C := \sum_{k=1}^{\infty} \|f_{n(k)} - f_{n(k+1)}\|_p < \infty$. Define $H_\infty = \lim_{N\to\infty} H_N$, where $H_N = \sum_{k=1}^{N} |f_{n(k)} - f_{n(k+1)}|$ for $1 \le N < \infty$. Use the triangle inequality to show that $\mu H_N^p \le C^p$ for all finite N. Then use Monotone Convergence to deduce that $\mu H_\infty^p \le C^p$.

(ii) Show that there exists a real-valued, measurable function f for which $f_{n(k)}(x) \to f(x)$ as $k \to \infty$, a.e. $[\mu]$.

(iii) Show that $|f_{n(k)} - f| \le \sum_{i=k}^{\infty} |f_{n(i)} - f_{n(i+1)}| \le H_\infty$ a.e. $[\mu]$. Use Dominated Convergence to deduce that $\|f_{n(k)} - f\|_p \to 0$ as $k \to \infty$.

(iv) Deduce from (iii) that f belongs to $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$ and $\|f_n - f\|_p \to 0$ as $n \to \infty$.

[20] For each random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ define

$$\|X\|_\infty := \inf\{c \in [0, \infty] : |X| \le c \text{ almost surely}\}.$$

Let $L^\infty := L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ denote the set of equivalence classes of real-valued random variables with $\|X\|_\infty < \infty$. Show that $\|\cdot\|_\infty$ is a norm on L^∞, which is a vector space, complete under the metric defined by $\|X\|_\infty$.

[21] Let $\{X_t : t \in T\}$ be a collection of $\overline{\mathbb{R}}$-valued random variables with possibly uncountable index set T. Complete the following argument to show that there exists a countable subset T_0 of T such that the random variable $X = \sup_{t\in T_0} X_t$ has the properties

(a) $X \ge X_t$ almost surely, for each $t \in T$

(b) if $Y \ge X_t$ almost surely, for each $t \in T$, then $Y \ge X$ almost surely

(The random variable X is called the ***essential supremum*** of the family. It is denoted by $\operatorname{ess\,sup}_{t\in T} X_t$. Part (b) shows that it is, unique up to an almost sure equivalence.)

(i) Show that properties (a) and (b) are unaffected by a monotone, one-to-one transformation such as $x \mapsto x/(1 + |x|)$. Deduce that there is no loss of generality in assuming $|X_t| \le 1$ for all t.

(ii) Let $\delta = \sup\{\mathbb{P} \sup_{t\in S} X_t : \text{ countable } S \subseteq T\}$. Choose countable T_n such that $\mathbb{P} \sup_{t\in T_n} X_t \ge \delta - 1/n$. Let $T_0 = \cup_n T_n$. Show that $\mathbb{P} \sup_{t\in T_0} X_t = \delta$.

(iii) Suppose $t \notin T_0$. From the inequality $\delta \ge \mathbb{P}(X_t \vee X) \ge \mathbb{P}X = \delta$ deduce that $X \ge X_t$ almost surely.

(iv) For a Y as in assertion (b), show that $Y \ge \sup_{t\in T_0} X_t = X$ almost surely.

[22] Let Ψ be a convex, increasing function for which $\Psi(0) = 0$ and $\Psi(x) \to \infty$ as $x \to \infty$. (For example, $\Psi(x)$ could equal x^p for some fixed $p \ge 1$, or $\exp(x) - 1$ or $\exp(x^2) - 1$.) Define $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ to be the set of all real-valued measurable functions on $\mathcal{X}$ for which $\mu\Psi(|f|/c_0) < \infty$ for some positive real c_0. Define

$\|f\|_\Psi := \inf\{c > 0 : \mu\Psi(|f|/c) \le 1\}$, with the convention that the infimum of an empty set equals $+\infty$. For each f, g in $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ and each real t prove the following assertions.

(i) $\|f\|_\Psi < \infty$. Hint: Apply Dominated Convergence to $\mu\Psi(|f|/c)$.

(ii) $f+g \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ and the triangle inequality holds: $\|f+g\|_\Psi \le \|f\|_\Psi + \|g\|_\Psi$. Hint: If $c > \|f\|_\Psi$ and $d > \|g\|_\Psi$, deduce that

$$\Psi\left(\frac{|f+g|}{c+d}\right) \le \frac{c}{c+d}\Psi\left(\frac{|f|}{c}\right) + \frac{d}{c+d}\Psi\left(\frac{|g|}{d}\right),$$

by convexity of Ψ.

(iii) $tf \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ and $\|tf\|_\Psi = |t|\,\|f\|_\Psi$.

REMARK. $\|\cdot\|_\Psi$ is called an Orlicz "norm"—to make it a true norm one should work with equivalence classes of functions equal μ almost everywhere. The L^p norms correspond to the special case $\Psi(x) = x^p$, for some $p \ge 1$.

[23] Define $\|f\|_\Psi$ and $\mathcal{L}^\Psi$ as in Problem [22]. Let $\{f_n\}$ be a Cauchy sequence in $\mathcal{L}^\Psi(\mu)$, that is, $\|f_n - f_m\|_\Psi \to 0$ as $\min(m,n) \to \infty$. Show that there exists an f in $\mathcal{L}^\Psi(\mu)$ for which $\|f_n - f\|_\Psi \to 0$, by following these steps.

(i) Let $\{g_i\}$ be a nonnegative sequence in $\mathcal{L}^\Psi(\mu)$ for which $C := \sum_i \|g_i\|_\Psi < \infty$. Show that the function $G := \sum_i g_i$ is finite almost everywhere and $\|G\|_\Psi \le \sum_i \|g_i\|_\Psi < \infty$. Hint: Use Problem [22] to show that $\mathbb{P}\Psi\left(\sum_{i\le n} g_i/C\right) \le 1$ for each n, then justify a passage to the limit.

(ii) Find an increasing sequence $\{n(k)\}$ such that $\sum_{k=1}^\infty \|f_{n(k)} - f_{n(k+1)}\|_\Psi < \infty$. Deduce that the functions $H_L := \sum_{k=L}^\infty |f_{n(k)} - f_{n(k+1)}|$ satisfy

$$\infty > \|H_1\|_\Psi \ge \|H_2\|_\Psi \ge \ldots \to 0.$$

(iii) Show that there exists a real-valued, measurable function f for which

$$|f_{n(k)}(x) - f(x)| \to 0 \qquad \text{as } k \to \infty, \text{ for } \mu \text{ almost all } x.$$

(iv) Given $\epsilon > 0$, choose L so that $\|H_L\|_\Psi < \epsilon$. For $i > L$, show that

$$\Psi\left(H_L/\epsilon\right) \ge \Psi\left(|f_{n(L)} - f_{n(i)}|/\epsilon\right) \to \Psi\left(|f_{n(L)} - f|/\epsilon\right).$$

Deduce that $\|f_{n(L)} - f\|_\Psi \le \epsilon$.

(v) Show that f belongs to $\mathcal{L}^\Psi(\mu)$ and $\|f_n - f\|_\Psi \to 0$ as $n \to \infty$.

[24] Let Ψ be a convex increasing function with $\Psi(0) = 0$, as in Problem [22]. Let Ψ^{-1} denote its inverse function. If $X_1, \ldots, X_N \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$, show that

$$\mathbb{P}\max_{i\le N}|X_i| \le \Psi^{-1}(N)\max_{i\le N}\|X_i\|_\Psi.$$

Hint: Consider $\Psi(\mathbb{P}\max|X_I|/C)$ with $C > \max_{i\le N}\|X_i\|_\Psi$.

REMARK. Compare with van der Vaart & Wellner (1996, page 96): if also $\limsup_{x,y\to\infty}\Psi(x)\Psi(y)/\Psi(cxy) < \infty$ for some constant $c > 0$ then $\|\max_{i\le N}|X_i|\|_\Psi \le K\Psi^{-1}(N)\max_{i\le N}\|X_i\|_\Psi$ for a constant K depending only on Ψ. See page 105 of their Problems and Complements for related counterexamples.

[25] For each θ in $[0,1]$ let $X_{n,\theta}$ be a random variable with a Binomial(n,θ) distribution. That is, $\mathbb{P}\{X_{n,\theta}=k\}=\binom{n}{k}\theta^k(1-\theta)^{n-k}$ for $k=0,1,\ldots,n$. You may assume these elementary facts: $\mathbb{P}X_{n,\theta}=n\theta$ and $\mathbb{P}(X_{n,\theta}-n\theta)^2=n\theta(1-\theta)$. Let f be a continuous function defined on $[0,1]$.

(i) Show that $p_n(\theta)=\mathbb{P}f(X_{n,\theta}/n)$ is a polynomial in θ.

(ii) Suppose $|f|\le M$, for a constant M. For a fixed ϵ, invoke (uniform) continuity to find a $\delta>0$ such that $|f(s)-f(t)|\le\epsilon$ whenever $|s-t|\le\delta$, for all s, t in $[0,1]$. Show that

$$|f(x/n)-f(\theta)|\le\epsilon+2M\{|(x/n)-\theta|>\delta\}\le\epsilon+\frac{2M|(x/n)-\theta|^2}{\delta^2}.$$

(iii) Deduce that $\sup_{0\le\theta\le1}|p_n(\theta)-f(\theta)|<2\epsilon$ for n large enough. That is, deduce that $f(\cdot)$ can be uniformly approximated by polynomials over the range $[0,1]$, a result known as the ***Weierstrass approximation theorem***.

[26] Extend the approximation result from Example <46> to the case of an infinite measure μ on $\mathcal{B}(\mathbb{R}^k)$ that gives finite measure to each compact set. Hint: Let B be a closed ball of radius large enough to ensure $\mu|f|B^c<\epsilon$. Write μ_B for the restriction of μ to B. Invoke the result from the Example to find a g in $\mathbb{C}_0$ such that $\mu_B|f-g|<\epsilon$. Find $\mathbb{C}_0$ functions $1\ge h_i\downarrow B$. Consider approximations gh_i for i large enough.

13. Notes

I recommend Royden (1968) as a good source for measure theory. The books of Ash (1972) and Dudley (1989) are also excellent references, for both measure theory and probability. Dudley's book contains particularly interesting historical notes.

See Hawkins (1979, Chapter 4) to appreciate the subtlety of the idea of a negligible set.

The result from Problem [10] is often attributed to (Pratt 1960), but, as he noted (in his 1966 Acknowledgment of Priority), it is actually much older.

Theorem <38> (the π–λ theorem for generating classes of sets) is often attributed to Dynkin (1960, Section 1.1), although Sierpiński (1928) had earlier proved a slightly stronger result (covering generation of sigma-rings, not just sigma-fields). I adapted the analogous result for classes of functions, Theorem <45>, from Protter (1990, page 7) and Dellacherie & Meyer (1978, page 14). Compare with the "Sierpiński Stability Lemma" for sets, and the "Functional Sierpiński Lemma" presented by Hoffmann-Jørgensen (1994, pages 8, 54, 60).

References

Ash, R. B. (1972), *Real Analysis and Probability*, Academic Press, New York.

Dellacherie, C. & Meyer, P. A. (1978), *Probabilities and Potential*, North-Holland, Amsterdam.

Dudley, R. M. (1989), *Real Analysis and Probability*, Wadsworth, Belmont, Calif.

Dynkin, E. B. (1960), *Theory of Markov Processes*, Pergamon.

Hawkins, T. (1979), *Lebesgue's Theory of Integration: Its Origins and Development*, second edn, Chelsea, New York.

Hoffmann-Jørgensen, J. (1994), *Probability with a View toward Statistics*, Vol. 1, Chapman and Hall, New York.

Oxtoby, J. (1971), *Measure and Category*, Springer-Verlag.

Pratt, J. W. (1960), 'On interchanging limits and integrals', *Annals of Mathematical Statistics* **31**, 74–77. Acknowledgement of priority, *same journal*, vol 37 (1966), page 1407.

Protter, P. (1990), *Stochastic Integration and Differential Equations*, Springer, New York.

Royden, H. L. (1968), *Real Analysis*, second edn, Macmillan, New York.

Sierpiński, W. (1928), 'Un théorème général sur les familles d'ensembles', *Fundamenta Mathematicae* **12**, 206–210.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Process: With Applications to Statistics*, Springer-Verlag.