Lectures on Integration

William G. Faris

March 4, 2001

Contents

Chapter 1

The integral: properties

1.1 Measurable functions

A *σ-algebra of measurable functions* $\mathcal{F}$ is a set of real functions defined on a set X. It must satisfy the following properties:

1. $\mathcal{F}$ is a vector space of functions.

2. $\mathcal{F}$ is a lattice of functions.

3. $\mathcal{F}$ is closed under pointwise monotone convergence.

4. $\mathcal{F}$ contains the constant functions.

To say that $\mathcal{F}$ is a vector space of functions is to say that f in $\mathcal{F}$ and $g \in \mathcal{F}$ imply that $f + g$ is in $\mathcal{F}$, and that a in $\mathbf{R}$ and f in $\mathcal{F}$ imply that af is in $\mathcal{F}$.

To say that $\mathcal{F}$ is a lattice of functions is to say that f in $\mathcal{F}$ and g in $\mathcal{F}$ imply that the minimum $f \wedge g$ is in $\mathcal{F}$ and that the maximum $f \vee g$ is in $\mathcal{F}$.

Notice that it follows that if f is in $\mathcal{F}$, then the absolute value given by the formula $|f| = (f \vee 0) - (f \wedge 0)$ is in $\mathcal{F}$.

To say that $\mathcal{F}$ is closed under pointwise upward convergence is to say that if each f_n is in $\mathcal{F}$ and if $f_n \uparrow f$, then f is in $\mathcal{F}$. To say that $f_n \uparrow f$ is to say that each $f_n \leq f_{n+1}$ pointwise and that f_n converges to f pointwise.

If $\mathcal{F}$ is closed under pointwise upward convergence, then it follows easily that $\mathcal{F}$ is closed under downward monotone convergence: If each f_n is in $\mathcal{F}$ and if $f_n \downarrow f$, then f is in $\mathcal{F}$. One just needs to apply the upward convergence property to the negatives of the functions.

To say that $\mathcal{F}$ is closed under pointwise monotone convergence is to say that it is closed under pointwise upward convergence and under pointwise downward convergence.

Let a be a real number and let f be a function. Define the function $1_{f>a}$ to have the value 1 at all points where $f > a$ and to have the value 0 at all points where $f \leq a$.

Theorem 1.1 *If f is in $\mathcal{F}$ and a is real, then $1_{f>a}$ is in $\mathcal{F}$.*

Proof: The function $f - f \wedge a$ is in $\mathcal{F}$. It is strictly greater than zero at precisely the points where $f > a$. The sequence of functions $g_n = n(f - f \wedge a)$ satisfies $g_n \uparrow \infty$ for points where $f > a$ and $g_n = 0$ at all other points. Hence the family of functions $g_n \wedge 1$ increases pointwise to $1_{f>a}$.

Theorem 1.2 *If f is a function, and if for each real number a the function $1_{f>a}$ is in $\mathcal{F}$, then f is in $\mathcal{F}$.*

Proof: First note that for $a < b$ the function $1_{a<f\leq b} = 1_{f>a} - 1_{f>b}$ is also in $\mathcal{F}$. Next, note that

$$f_n = \sum_{k=-\infty}^{\infty} \frac{k}{2^n} 1_{\frac{k}{2^n} < f \leq \frac{k+1}{2^n}} \tag{1.1}$$

is in $\mathcal{F}$. However $f_n \uparrow f$ as $n \to \infty$.

Theorem 1.3 *If f is in $\mathcal{F}$, then so is f^2.*

Proof: For $a \geq 0$ the condition $f^2 > a$ is equivalent to the condition $|f| > \sqrt{a}$.

Theorem 1.4 *If f, g are in $\mathcal{F}$, then so is the pointwise product $f \cdot g$.*

Proof: $(f+g)^2 - (f-g)^2 = 4f \cdot g$.

This last theorem shows that $\mathcal{F}$ is not only closed under addition, but also under multiplication. Thus $\mathcal{F}$ deserves to be called an algebra. It is called a σ algebra because of the closure under monotone limits. We shall now see that there is actually closure under all pointwise limits.

Theorem 1.5 *Let f_n be in $\mathcal{F}$ and $f_n \to f$ pointwise. Then f is in $\mathcal{F}$.*

Proof: Let $n < m$ and let $h_{nm} = f_n \wedge f_{n+1} \wedge \cdots \wedge f_m$. Then $h_{nm} \downarrow h_n$ as $m \to \infty$, where h_n is the infimum of the f_k for $k \geq n$. However $h_n \uparrow f$.

The trick in this proof is to write a general limit as an increasing limit followed by a decreasing limit. We shall see in the following that this is a very important idea in integration.

If we are given a set of functions, then the σ-algebra of functions generated by this set is the smallest σ-algebra of functions that contains the original set. The *Borel* σ-algebra of functions is the σ-algebra of functions $\mathcal{B}$ on $\mathbf{R}^n$ generated by the coordinates $x_1, \ldots, x_n$. The following theorem shows that measurable functions are closed under nonlinear operations in a very strong sense.

Theorem 1.6 *Let $f_1, \ldots, f_n$ be in a σ-algebra $\mathcal{F}$ of measurable functions. Let ϕ be a Borel function. Then $\phi(f_1, \ldots, f_n)$ is also in $\mathcal{F}$.*

Proof: Let $\mathcal{B}'$ be the set of all functions ϕ such that $\phi(f_1, \ldots, f_n)$ is in $\mathcal{F}$. It is easy to check that $\mathcal{B}'$ is a σ-algebra of functions. Since $f_1, \ldots, f_n$ are each measurable, it follows that the coordinate functions $x_1, \ldots, x_n$ belong to $\mathcal{B}'$. It follows from these two facts that $\mathcal{B}$ is a subset of $\mathcal{B}'$.

1.2 Integration

An *integral* is a function μ defined on the positive elements of a σ-algebra of measurable functions with values in the interval $[0, \infty]$. It must satisfy the following properties:

1. μ is additive and respects scalar multiplication by positive scalars.

2. μ satisfies the monotone convergence property.

The first property says that for $f \geq 0$ and $g \geq 0$ we always have $\mu(f + g) = \mu(f) + \mu(g)$. Furthermore, for $a \geq 0$ and $f \geq 0$ we have $\mu(af) = a\mu(f)$. In this context the product 0 times ∞ is defined to be 0.

Theorem 1.7 *If $0 \leq f \leq g$, then $0 \leq \mu(f) \leq \mu(g)$.*

Proof: Clearly $(g - f) + f = g$. So $\mu(g - f) + \mu(f) = \mu(g)$. But $\mu(g - f) \geq 0$.

The second property says that if each $f_n \geq 0$ and $f_n \uparrow f$ as $n \to \infty$, then $\mu(f_n) \to \mu(f)$ as $n \to \infty$. This is usually called the *monotone convergence theorem*.

If f is a measurable function, then its positive part $(f \vee 0) \geq 0$ and its negative part $-(f \wedge 0) \geq 0$. So they each have integrals. If at least one of these integrals is finite, then we may define the integral of f to be

$$\mu(f) = \mu(f \vee 0) - \mu(-(f \wedge 0)). \tag{1.2}$$

However, the expression $\infty - \infty$ is undefined!

Theorem 1.8 *If $\mu(|f|) < \infty$, then $\mu(f)$ is defined, and*

$$|\mu(f)| \leq \mu(|f|). \tag{1.3}$$

It is very common to denote

$$\mu(f) = \int f \, d\mu \tag{1.4}$$

or even

$$\mu(f) = \int f(x) \, d\mu(x). \tag{1.5}$$

This notation is suggestive in the case when there is more than one integral in play. Say that ν is an integral, and $w \geq 0$ is a measurable function. Then the integral $\mu(f) = \nu(fw)$ is defined. We would write this as

$$\int f(x) \, d\mu(x) = \int f(x) \, w(x) d\nu(x). \tag{1.6}$$

So the relation between the two integrals would be $d\mu(x) = w(x)d\nu(x)$. This suggests that $w(x)$ plays the role of a derivative of one integral with respect to the other.

1.3 Convergence theorems

Theorem 1.9 *(improved monotone convergence) If $\mu(f_1) > -\infty$ and $f_n \uparrow f$, then $\mu(f_n) \to \mu(f)$.*

Proof: Apply monotone convergence to $f_n - f_1$.

Theorem 1.10 *(Fatou) If each $f_n \geq 0$ and if $f_n \to f$ pointwise as $n \to \infty$, then $\mu(f) \leq \lim_{n\to\infty} \inf_{k\geq n} \mu(f_k)$.*

Proof: Let h_n be the infimum of the f_k for $k \geq n$. Then $h_n \leq f_k$ for each $k \geq n$, and so $\mu(h_n) \leq \mu(f_k)$ for each $k \geq n$. In particular, $\mu(h_n) \leq \inf_{k\geq n} \mu(f_k)$. However $h_n \uparrow f$. Both sides of this inequality are increasing, and so by monotone convergence $\mu(f) \leq \lim_{n\to\infty} \inf_{k\geq n} \mu(f_k)$.

Here is an important special case of Fatou's lemma. If each $f_n \geq 0$ and $0 \leq \mu(f_n) \leq M$ and $f_n \to f$ pointwise as $n \to \infty$, then $0 \leq \mu(f) \leq M$. In other words, one can lose mass, but not gain it.

Theorem 1.11 *(dominated convergence) Let $|f_n| \leq g$ for each n, where $\mu(g) < \infty$. Then if $f_n \to f$ as $n \to \infty$, then $\mu(f_n) \to \mu(f)$ as $n \to \infty$.*

Proof: Let u_n be the supremum of $k \geq n$ of the f_k, and let h_n be the infimum for $k \geq n$ of the f_k. Then $-g \leq h_n \leq f_n \leq u_n \leq g$. However $h_n \uparrow f$ and $u_n \downarrow f$. It follows from improved monotone convergence that $\mu(h_n) \to \mu(f)$ and $\mu(u_n) \to \mu(f)$. Since $\mu(h_n) \leq \mu(f_n) \leq \mu(u_n)$, it follows that $\mu(f_n) \to \mu(f)$.

Theorem 1.12 *(Tonelli for positive sums) If $w_k \geq 0$, then*

$$\mu(\sum_{k=1}^{\infty} w_k) = \sum_{k=1}^{\infty} \mu(w_k) \tag{1.7}$$

Proof: This theorem says that for positive functions integrals and sums may be interchanged. This is the monotone convergence theorem in disguise.

Theorem 1.13 *(Fubini for absolutely convergent sums) If $\mu(\sum_k |w_k|) < \infty$, then*

$$\mu(\sum_{k=1}^{\infty} w_k) = \sum_{k=1}^{\infty} \mu(w_k). \tag{1.8}$$

Proof: This theorem says that absolute convergence implies that integrals and sums may be interchanged. Since

$$|\sum_{k=1}^{n} w_k| \leq \sum_{k=1}^{n} |w_k| \leq \sum_{k=1}^{\infty} |w_k|, \tag{1.9}$$

this follows from the dominated convergence theorem.

1.4 Measure

If E is a subset of X, then 1_E is the indicator function of E. Its value is 1 for every point in E and 0 for every point not in E. The set E is said to be measurable if the function 1_E is measurable. The *measure* of such an E is $\mu(1_E)$. This is often denoted $\mu(E)$.

Sometimes we denote a subset of X by a condition that defines the subset. Thus, for instance, the set of all points where $f < a$ is denoted $f < a$, and its measure is $\mu(f < a)$. This is more economical than writing this explicitly as in integral $\mu(1_{f<a})$.

Theorem 1.14 *If the set where $f \neq 0$ has measure zero, then $\mu(|f|) = 0$.*

Proof: For each n the function $|f| \wedge n \leq n1_{|f|>0}$ and so has integral $\mu(|f| \wedge n) \leq n \cdot 0 = 0$. However $|f| \wedge n \uparrow |f|$ as $n \to \infty$. So from monotone convergence $\mu(|f|) = 0$.

The preceding theorem shows that changing a function on a set of measure zero does not change its integral. Thus, for instance, if we change g to $h = g + f$, then $|\mu(h) - \mu(g)| = |\mu(f)| \leq \mu(|f|) = 0$.

There is a terminology that is standard in this situation. If a property of points is true except on a measure zero, then it is said to hold *almost everywhere*. Thus the theorem would be stated as saying that if $f = 0$ almost everywhere, then its integral is zero. Similarly, if $g = h$ almost everywhere, then g and h have the same integral.

Theorem 1.15 *(Chebyshev inequality) Let f be a measurable function and let a be a real number in the range of f. Let ϕ be a function defined on the range of f such that $x_1 \leq x_2$ implies $0 \leq \phi(x_1) \leq \phi(x_2)$. Then*

$$\phi(a)\mu(f \geq a) \leq \mu(\phi(f)). \tag{1.10}$$

Proof: This follows from the pointwise inequality

$$\phi(a)1_{f \geq a} \leq \phi(a)1_{\phi(f) \geq \phi(a)} \leq \phi(f). \tag{1.11}$$

The Chebyshev inequality is often used in the case when $f \geq 0$ and $\phi(x) = x$ or $\phi(x) = x^2$. Another important case is $\phi(x) = e^{tx}$ with $t > 0$. Here we will use it in the simple form $a\mu(\geq a) \leq \mu(f)$ for $f \geq 0$.

Theorem 1.16 *If $\mu(|f|) = 0$, then the set where $f \neq 0$ has measure zero.*

Proof: By the Chebyshev inequality, for each n we have $\mu(1_{|f|>1/n} \leq n\mu(|f|) = n \cdot 0 = 0$. However as $n \to \infty$, the functions $1_{|f|>1/n} \uparrow 1_{|f|>0}$. So $\mu(1_{|f|>0}) = 0$.

The above theorem also has a statement in terms of an almost everywhere property. It says that if $|f|$ has integral zero, then $f = 0$ almost everywhere.

Theorem 1.17 *Let $f_n \uparrow$ be an increasing sequence of positive measurable functions such that for some $M < \infty$ and for all n the integrals $\mu(f_n) \leq M$. Let E be the set on which $f_n \to \infty$. Then $\mu(E) = 0$.*

Proof: Let f be the limit of the f_n; the function f can assume the value ∞. For each real $a > 0$ the functions $f_n \wedge a$ increase to a limit $f \wedge a$. Since $\mu(f_n \wedge a) \leq \mu(f_n) \leq M$, it follows that $\mu(f \wedge a) \leq M$. Let E_a be the set on $f \wedge a \geq a$. By Chebyshev's inequality $a\mu(E_a) \leq \mu(f \wedge a) \leq M$. Since E is a subset of E_a, it follows that $a\mu(E) \leq M$. This can happen for all real $a > 0$ only when $\mu(E) = 0$.

Again this result has an almost everywhere statement. It says that if an increasing sequence of positive functions has a finite upper bound on its integral, then the limit of the functions is finite almost everywhere.

Theorem 1.18 *An integral is uniquely determined by the corresponding measure.*

Let $f \geq 0$ be a measurable function. Define

$$f_n = \sum_{k=0}^{\infty} \frac{k}{2^n} 1_{\frac{k}{2^n} < f \leq \frac{k+1}{2^n}}. \tag{1.12}$$

The integral of f_n is determined by the measures of the sets where $\frac{k}{2^n} < f \leq \frac{k+1}{2^n}$. However $f_n \uparrow f$, and so the integral of f is determined by the corresponding measure.

This theorem justifies a certain amount of confusion between the notion of measure and the notion of integral. In fact, this whole subject is sometimes called measure theory.

1.5 Fubini's theorem

If X, Y are sets, their *Cartesian* product $X \times Y$ is the set of all ordered pairs (x, y) with x in X and y in Y.

If f is a function on X and g is a function on Y, then their *tensor product* $f \otimes g$ is the function on the Cartesian product $X \times Y$ given by

$$(f \otimes g)(x, y) = f(x)g(y). \tag{1.13}$$

Sometimes such functions are called *decomposable*. Another term is *separable*, as in the expression "separation of variables." If $\mathcal{F}$ is a σ-algebra of measurable functions on X and $\mathcal{G}$ is a σ-algebra of measurable functions on Y, then $\mathcal{F} \otimes \mathcal{G}$ is the *product σ-algebra* generated by the functions $f \otimes g$ with f in $\mathcal{F}$ and g in $\mathcal{G}$.

We say that an integral μ is *σ-finite* if there is a sequence of functions $f_n \geq 0$ with $\mu(f_n) < \infty$ for each n and with $f_n \uparrow 1$.

Theorem 1.19 *Let $\mathcal{F}$ consist of measurable functions on X and $\mathcal{G}$ consist of measurable functions on Y. Then $\mathcal{F} \otimes \mathcal{G}$ consists of measurable functions on $X \times Y$. If μ is a σ-finite integral on $\mathcal{F}$ and ν is a σ-finite integral on $\mathcal{G}$, then*

there exists a unique integral $\mu \times \nu$ on $\mathcal{F} \otimes \mathcal{G}$ with the property that for every pair $f \geq 0$ in $\mathcal{F}$ and $g \geq 0$ in $\mathcal{G}$ we have

$$(\mu \times \nu)(f \otimes g) = \mu(f)\nu(g). \tag{1.14}$$

The product of integrals in this theorem may be of the form $0 \cdot \infty$ or $\infty \cdot 0$. In that case we use the convention that is standard in integration theory: $0 \cdot \infty = \infty \cdot 0 = 0$. The integral described in the above theorem is called the *product integral*. The corresponding measure is called the *product measure*. The defining property may also be written in the more explicit form

$$\int f(x)g(y)\, d(\mu \times \nu)(x, y) = \int f(x)\, d\mu(x) \int g(y)\, d\nu(y). \tag{1.15}$$

The definition of product integral does not immediately give a useful way to compute the integral of functions that do not compose into a product. For this we need Tonelli's theorem and Fubini's theorem. Here is another definition. Let ν be a measure defined on functions of y in Y. Let h be a function on $X \times Y$. Then the partial integral $\nu(h \mid X)$ is the function on X defined for each x as the ν integral of the function $y \mapsto h(x, y)$. It should be thought of the integral of $h(x, y)$ over y with the x coordinate fixed. Similarly, in μ is a measure defined on functions of y in Y, then the partial integral $\mu(h \mid Y)$ is defined in the analogous way. It is the integral of $h(x, y)$ over x with the y coordinate fixed. These notations are not standard, but they are suggested by a somewhat similar notation for conditional expectation in probability theory.

Theorem 1.20 *(Tonelli) Let μ and ν be σ-finite integrals defined for measurable functions on X and Y respectively. Let $h \geq 0$ be a measurable function on $X \times Y$. Then*

$$(\mu \times \nu)(h) = \mu(\nu(h \mid X)) = \nu(\mu(h \mid Y)). \tag{1.16}$$

The identity in Tonelli's theorem may of course also be written as saying that

$$\int h(x, y)\, d(\mu \times \nu)(x, y) = \int [\int h(x, y)\, d\nu(y)]\, d\mu(x) = \int [\int h(x, y)\, d\mu(x)]\, d\nu(y) \tag{1.17}$$

whenever $h(x, y) \geq 0$. There is a technicality about the integrals that occur in Tonelli's theorem. It is quite possible that even if the function h that one starts with has values in the interval $[0, \infty)$, then a partial integral such as $\nu(h \mid X)$ is a function with values in the interval $[0, \infty]$. So it is convenient to extend the notion of measurable function to positive functions that can assume the value ∞. We define the integral for such positive measurable functions in such a way that the monotone convergent theory is satisfied.

The proof of Tonelli's theorem may be reduced to the theorem about the existence of the product integral. (Of course this theorem was not proved above.) The idea is to show first that it is true for decomposible functions and then extend to all measurable functions.

Theorem 1.21 *(Fubini) Let μ and ν be σ-finite integrals defined for measurable functions on X and Y respectively. Let h be a measurable function on $X \times Y$ such that $(\mu \times \nu)(|h|) < \infty$. Thus $(\mu \times \nu)(h)$ is defined. Then $\nu(|h| \mid X) < \infty$ μ-almost everywhere, and $\mu(|h| \mid Y)) < \infty$ ν-almost everywhere. Thus $\nu(h \mid X)$ is defined μ-almost everywhere, and $\mu(h \mid Y))$ is defined ν-almost everywhere. Furthermore $\mu(|\nu(h \mid X)|) < \infty$ and $\nu(|\mu(h \mid Y)|) < \infty$. Thus $\mu(\nu(h \mid X))$ and $\nu(\mu(h \mid Y))$ are defined. Finally,*

$$(\mu \times \nu)(h) = \mu(\nu(h \mid X)) = \nu(\mu(h \mid Y)). \tag{1.18}$$

The identity in Fubini's theorem may of course also be written as saying that

$$\int h(x,y)\, d(\mu \times \nu)(x,y) = \int [\int h(x,y)\, d\nu(y)]\, d\mu(x) = \int [\int h(x,y)\, d\mu(x)]\, d\nu(y) \tag{1.19}$$

whenever $|h(x,y)|$ has a finite integral with respect to the product measure. There is also a technicality about the integrals that occur in Fubini's theorem. It is quite possible that the partial integral $\nu(h \mid X)$ is undefined at certain points, due to an $\infty - \infty$ problem. One could worry about whether it is meaningful to integrate such a function. However under the hypotheses of Fubini's theorem, such a problem would occur only on a set of μ measure zero. So it is convenient to assume that we have defined the integral for certain functions that are defined almost everywhere and whose absolute values have finite integrals.

The proof of Fubini's theorem from Tonelli's theorem is not difficult. One merely applies Tonelli's theorem to the positive and negative parts of the function.

Tonelli's theorem and Fubini's theorem are often used together to justify an interchange of order of integration. Say that one can show that

$$\int [\int |h(x,y)|\, d\nu(y)]\, d\mu(x) < \infty. \tag{1.20}$$

Then by Tonelli's theorem

$$\int |h(x,y)|\, d(\mu \times \nu)(x,y) < \infty. \tag{1.21}$$

However then it follows from Fubini's theorem that

$$\int [\int h(x,y)\, d\nu(y)]\, d\mu(x) = \int [\int h(x,y)\, d\mu(x)]\, d\nu(y). \tag{1.22}$$

Finally, a technical note. We have been assuming that the measure spaces are σ-finite. Why is this necessary? Here is an example. Let $\mu(f) = \int_0^1 f(x)\, dx$ be the usual uniform Lebesgue integral on the interval $[0,1]$. Let $\nu(g) = \sum_y g(y)$ be summation indexed by the points in the interval $[0,1]$. The measure ν is not σ-finite, since there are uncountably many points in $[0,1]$. Finally, let $h(x,y) = 1$ if $x = y$, and $h(x,y) = 0$ for $x \neq y$. Now for each x, the sum

$\sum_y h(x,y) = 1$. So the integral over x is also 1. On the other hand, for each y the integral $\int h(x,y)\,dx = 0$, since the integrand is zero except for a single point of μ measure zero. So the sum over y is zero. Thus the two orders of integration give different results. Tonelli's theorem does not work. Allowing measure spaces that are not σ-finite is like allowing uncountable sums. The theory of integration only works well for countable sums.

1.6 Problems

1. Let $\mathcal{B}$ be the smallest σ-algebra of real functions on $\mathbf{R}$ containing the function x. This is called the σ-algebra of Borel functions. Show that every continuous function is a Borel function.

2. Show that every monotone function is a Borel function.

3. Can a Borel function be discontinuous at every point?

   The preceding problems show that it is difficult to think of a specific example of a real function on $\mathbf{R}$ that is not a Borel function. However it may be proved that they exist, and examples are known.

   The following problems deal with certain smaller σ-algebras of functions. Such smaller σ-algebras are very important in probability theory. For instance, consider the σ-algebra generated by some function f. The idea is that it represents all functions that can be computed starting from information about the value of f. Thus if f has somehow been measured in an experiment, then all the other functions in the σ-algebra are also measurable in the same experiment.

4. Let $\mathcal{B}'$ be the smallest σ-algebra of functions on $\mathbf{R}$ containing the function x^2. Show that $\mathcal{B}'$ is not equal to $\mathcal{B}$. Which algebra of measurable functions is bigger (that is, which one is a subset of the other)?

5. Consider the σ-algebras of functions generated by $\cos(x)$, $\cos^2(x)$, and $\cos^4(x)$. Compare them with the σ-algebras in the previous problem and with each other. (Thus specify which ones are subsets of other ones.)

6. This problem is to show that one can get convergence theorems when the family of functions is indexed by real numbers. Prove that if $f_t \to f$ pointwise as $t \to t_0$, $|f_t| \le g$ pointwise, and $\mu(g) < \infty$, then $\mu(f_t) \to \mu(f)$ as $t \to t_0$.

   The usual Lebesgue integral $\lambda(f) = \int_{-\infty}^{\infty} f(x)\,dx$ is defined for positive Borel functions. It then extends to Borel functions that have either a positive or a negative part with finite integral.

7. Show that if f is a Borel function and $\int_{-\infty}^{\infty} |f(x)|\,dx < \infty$, then $F(b) = \int_{-\infty}^{b} f(x)\,dx$ is continuous.

8. Must the function F in the preceding problem be differentiable at every point? Discuss.

9. Show that
$$\int_0^\infty \frac{\sin(e^x)}{1+nx^2}\,dx \to 0 \tag{1.23}$$
as $n \to \infty$.

10. Show that
$$\int_0^1 \frac{n\cos(x)}{1+n^2x^{\frac{3}{2}}}\,dx \to 0 \tag{1.24}$$
as $n \to \infty$.

11. Evaluate
$$\lim_{n\to\infty} \int_a^\infty \frac{n}{1+n^2x^2}\,dx \tag{1.25}$$
as a function of a.

12. Consider the integral
$$\int_{-\infty}^\infty \frac{1}{\sqrt{1+nx^2}}\,dx. \tag{1.26}$$
Show that the integrand is monotone decreasing and converges pointwise as $n \to \infty$, but the integral of the limit is not equal to the limit of the integrals. How does this relate to the monotone convergence theorem?

13. Let $\delta(x)$ be a Borel function with
$$\int_{-\infty}^\infty |\delta(x)|\,dx < \infty \tag{1.27}$$
and
$$\int_{-\infty}^\infty \delta(x)\,dx = 1 \tag{1.28}$$
Let
$$\delta_\epsilon(x) = \delta(\frac{x}{\epsilon})\frac{1}{\epsilon}. \tag{1.29}$$
Let g be bounded and continuous. Show that
$$\int_{-\infty}^\infty \delta_\epsilon(y)g(y)\,dy \to g(0) \tag{1.30}$$
as $\epsilon \to 0$. This problem gives a very general class of functions $\delta_\epsilon(x)$ such that integration with $\delta_\epsilon(x)\,dx$ converges to the Dirac delta integral δ given by $\delta(g) = g(0)$.

14. Let f be bounded and continuous. Show for each x the convolution
$$\int_{-\infty}^\infty \delta_\epsilon(x-z)f(z)\,dz \to f(x) \tag{1.31}$$
as $\epsilon \to 0$.

15. Prove *countable subadditivity*:

$$\mu(\bigcup_{n=1}^{\infty} A_n) \leq \int (\sum_{n=1}^{\infty} 1_{A_n})\, d\mu = \sum_{n=1}^{\infty} \mu(A_n). \tag{1.32}$$

Show that if the A_n are disjoint this is an equality (countable additivity).

16. Recall that $f_n(x) \to f(x)$ as $n \to \infty$ means $\forall \epsilon > 0\, \exists N\, \forall n \geq N\, |f_n(x) - f(x)| < \epsilon$. Show that $f_n \to f$ almost everywhere is equivalent to

$$\mu(\{x \mid \exists \epsilon > 0\, \forall N\, \exists n \geq N\, |f_n(x) - f(x)| \geq \epsilon\}) = 0. \tag{1.33}$$

17. Show that $f_n \to f$ almost everywhere is equivalent to for all $\epsilon > 0$

$$\mu(\{x \mid \forall N\, \exists n \geq N\, |f_n(x) - f(x)| \geq \epsilon\}) = 0. \tag{1.34}$$

18. Suppose that the measure of the space is finite. Show that $f_n \to f$ almost everywhere is equivalent to for all $\epsilon > 0$

$$\lim_{N \to \infty} \mu(\{x \mid \exists n \geq N\, |f_n(x) - f(x)| \geq \epsilon\}) = 0. \tag{1.35}$$

Show that this is not equivalent in the case when the measure of the space may be infinite.

Note: Convergence almost everywhere occurs in the strong law of large numbers.

19. Say that $f_n \to f$ *in measure* if for all $\epsilon > 0$

$$\lim_{N \to \infty} \mu(\{x \mid |f_N(x) - f(x)| \geq \epsilon\}) = 0. \tag{1.36}$$

Show that if the measure of the space is finite, then $f_n \to f$ almost everywhere implies $f_n \to f$ in measure. Note: Convergence in measure occurs in the weak law of large numbers.

Chapter 2

Function spaces

2.1 Spaces of continuous functions

This section records notations for spaces of functions. Ordinarily we think of real or of complex functions. Which one depends on context.

The space $C(X)$ consists of all continuous functions. The space $B(X)$ consists of all bounded functions. It is a Banach space in a natural way. The space $BC(X)$ consists of all bounded continuous functions. It is a somewhat smaller Banach space.

Now look at the special case when X is a metric space such that each point has a compact neighborhood. (More generally, X could be a Hausdorff topological space instead of a metric space.) For example X could be an open subset of $\mathbf{R}^n$. The space $C_c(X)$ consists of all continuous functions, each one of which has compact support. The space $C_0(X)$ is the closure of $C_c(X)$ in $BC(X)$. It is itself a Banach space. It is the space of continuous functions that vanish at infinity.

The relation between these spaces is that $C_c(X) \subset C_0(X) \subset BC(X)$. They are all equal when X compact. When X is locally compact, then $C_0(X)$ is the best behaved.

In the following we shall need the concept of *dual space* of a Banach space. The dual space consists of all continuous linear functions from the Banach space to the scalars. It is itself a Banach space. For certain Banach spaces of functions the linear functionals in the dual space may be realized by integration with certain elements of a Banach space of functions or of measures. For example, $C_(X)$ is a Banach space, and its dual space $M(X)$ is a Banach space consisting of signed Radon measures of finite variation. (A signed measure of finite variation is the difference of two positive measures that each assigns finite measure to the space X. Radon measures will be discussed later.) If σ is in $M(X)$, then it defines the linear functional $f \mapsto \int f(x)\, d\sigma(x)$, and all elements of the dual space arise in this way.

2.2 L^p spaces

Fix a set X and a σ-algebra $\mathcal{F}$ of measurable functions. Let $0 < p < \infty$. Define

$$\|f\|_p = \mu(|f|^p)^{\frac{1}{p}}. \tag{2.1}$$

Define $L^p(\mu)$ to be the set of all f in $\mathcal{F}$ such that $\|f\|_p < \infty$.

Theorem 2.1 *For $0 < p < \infty$, the space $L^p(\mu)$ is a vector space.*

Proof: It is obvious that L^p is closed under scalar multiplication. The problem is to prove that it is closed under addition. However if f, g are each in L^p, then

$$|f+g|^p \le [2(|f| \vee |g|)]^p \le 2^p(|f|^p + |g|^p). \tag{2.2}$$

Thus $f + g$ is also in L^p.

The function x^p is increasing for every $p > 0$, but it is convex only for $p \ge 1$. This is the key to the following fundamental inequality.

Theorem 2.2 *(Minkowski's inequality) If $1 \le p < \infty$, then*

$$\|f+g\|_p \le \|f\|_p + \|g\|_p. \tag{2.3}$$

Proof: Let $c = \|f\|_p$ and $d = \|g\|_p$. Then by the fact that x^p is increasing and convex

$$\left|\frac{f+g}{c+d}\right|^p \le \left(\frac{c}{c+d}\left|\frac{f}{c}\right| + \frac{d}{c+d}\left|\frac{g}{d}\right|\right)^p \le \frac{c}{c+d}\left|\frac{f}{c}\right|^p + \frac{d}{c+d}\left|\frac{g}{d}\right|^p. \tag{2.4}$$

Integrate. This gives

$$\mu\left(\left|\frac{f+g}{c+d}\right|^p\right) \le 1. \tag{2.5}$$

Thus $\|f+g\|_p \le c + d$.

It is a fact that $\mu(|f|^p) = 0$ if and only if $f = 0$ almost everywhere. Thus in order to get a normed vector space, we need to define the space in such a way that every two functions g, h that differ by a function $f = h - g$ that is zero almost everywhere determine the same element of the vector space. With this convention, we have the property that $\|f\|_p = 0$ implies $f = 0$ as an element of this vector space.

Theorem 2.3 *(dominated convergence for L^p) Let $f_n \to f$ pointwise. Suppose that there is a $g \ge 0$ in L^p such that each $|f_n| \le g$. Then $f_n \to f$ in L^p, that is, $\|f_n - f\|_p \to 0$.*

Proof: If each $|f_n| \le g$ and $f_n \to f$ pointwise, then $|f| \le g$. Thus $|f_n - f| \le 2g$ and $|f_n - f|^p \le 2^p g^p$. Since g^p has finite integral, the integral of $|f_n - f|^p$ approaches zero, by the usual dominated convergence theorem.

It would be an error to think that just because $g_n \to g$ in the L^p sense it would follow that $g_n \to g$ almost everywhere. Being close on the average does not imply being close at a particular point.

Example: For each $n = 1, 2, 3, \ldots$, write $n = 2^k + j$, where $k = 0, 1, 2, 3, \ldots$ and $0 \leq j < 2^k$. Consider a sequence of functions defined on the unit interval $[0, 1]$ with the usual Lebesgue measure. Let $g_n = 1$ on the interval $[j/2^k, (j + 1)/2^k]$ and $g_n = 0$ elsewhere in the unit interval. Then the L^1 norm of g_n is $1/2^k$, so the $g_n \to 0$ in the L^1 sense. On the other hand, given x in $[0, 1]$, there are infinitely many n for which $g_n(x) = 0$ and there are infinitely many n for which $g_n(x) = 1$. So pointwise convergence fails at each point.

Lemma 2.1 *If a normed vector space is such that every absolutely convergent series is convergent, then the space is complete and hence is a Banach space.*

Proof: Suppose that g_n is a Cauchy sequence. This means that for every $\epsilon > 0$ there is an N such that $m, n \geq N$ implies $\|g_m - g_n\| < \epsilon$. The idea is to show that g_n has a subsequence that converges very rapidly. Let ϵ_k be a sequence such that $\sum_{k=1}^{\infty} \epsilon_k < \infty$. In particular, for each k there is an N_k such that $m, n \geq N_k$ implies $\|g_m - g_n\| < \epsilon_k$. The desired subsequence is the g_{N_k}. Define a sequence $f_1 = g_{N_1}$ and $f_j = g_{N_j} - g_{N_{j-1}}$ for $j \geq 2$. Then

$$g_{N_k} = \sum_{j=1}^{k} f_j. \tag{2.6}$$

Furthermore,

$$\sum_{j=1}^{\infty} \|f_j\| \leq \|f_1\| + \sum_{j=2}^{\infty} \epsilon_{j-1} < \infty. \tag{2.7}$$

Hence there exists a g such that the subsequence g_{N_k} converges to g. Since the sequence g_n is Cauchy, it also must converge to the same g.

Theorem 2.4 *For $1 \leq p < \infty$ the space $L^p(\mu)$ is a Banach space.*

Proof: Suppose that $\sum_{j=1}^{\infty} f_j$ is absolutely convergent in L^p, that is,

$$\sum_{j=1}^{\infty} \|f_j\|_p = B < \infty. \tag{2.8}$$

Then by using Minkowski's inequality

$$\|\sum_{j=1}^{k} |f_j|\|_p \leq \sum_{j=1}^{k} \|f_j\|_p \leq B. \tag{2.9}$$

By the monotone convergence theorem $h = \sum_{j=1}^{\infty} |f_j|$ is in L^p with L^p norm bounded by B. In particular, it is convergent almost everywhere. It follows that the series $\sum_{j=1}^{\infty} f_j$ converges almost everywhere to some limit g. The sequence

$\sum_{j=1}^{k} f_j$ is dominated by h in L^p and converges pointwise to $\sum_{j=1}^{\infty} f_j$. Therefore, by the dominated convergence theorem, it converges to the same limit g in the L^p norm.

Corollary 2.1 *If $1 \leq p < \infty$ and if $g_n \to g$ in the L^p norm sense, then there is a subsequence g_{N_k} such that g_{N_k} converges to g almost everywhere.*

Proof: Let $g_n \to g$ as $n \to \infty$ in the L^p sense. Then g_n is a Cauchy sequence in the L^p sense. Let ϵ_k be a sequence such that $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Let N_k be a subsequence such that $n \geq N_k$ implies $\|g_n - g_{N_k}\|_p < \epsilon_k$. Define a sequence f_k such that

$$g_{N_k} = \sum_{j=1}^{k} f_j. \tag{2.10}$$

Then $\|f_j\|_p = \|g_{N_j} - g_{N_{j-1}}\|_p \leq \epsilon_{j-1}$ for $j \geq 2$. By the monotone convergence theorem

$$h = \sum_{j=1}^{\infty} |f_j| \tag{2.11}$$

converges in L^p and is finite almost everywhere. It follows that

$$g = \sum_{j=1}^{\infty} f_j \tag{2.12}$$

converges in L^p and also converges almost everywhere In particular, $g_{N_k} \to g$ as $k \to \infty$ almost everywhere.

In order to complete the picture, define

$$\|f\|_\infty = \inf\{M \geq 0 \mid |f| \leq M \text{ almost everywhere }\}. \tag{2.13}$$

It is not hard to show that there is actually a least such M. The following theorem is also simple.

Theorem 2.5 *The space $L^\infty(\mu)$ is a Banach space.*

2.3 Duality of L^p spaces

Lemma 2.2 *Let $p > 1$ and $q > 1$ with $1/p + 1/q = 1$. If $x \geq 0$ and $y \geq 0$, then*

$$xy \leq \frac{1}{p}x^p + \frac{1}{q}y^q. \tag{2.14}$$

Proof: Fix y and consider the function $xy - \frac{1}{p}x^p$. Since $p > 1$ it rises to a maximum and then dips below zero. It has its maximum where the derivative is equal to zero, that is, where $y - x^{p-1} = 0$. At this point $x = y^{\frac{1}{p-1}} = y^{q-1}$. The maximum value is $y^q - \frac{1}{p}y^q = \frac{1}{q}y^q$.

The above result may also be written in the form

$$a^{\frac{1}{p}} b^{\frac{1}{q}} \leq \frac{1}{p} a + \frac{1}{q} b, \tag{2.15}$$

where the weights satisfy $1/p + 1/q = 1$. In this form it is called the inequality of the geometric and arithmetic mean. The original version of this inequality, of course, is the simple but useful $\sqrt{ab} \leq (a+b)/2$.

Theorem 2.6 *(Hölder's inequality) Suppose that* $1 < p < \infty$ *and that* $1/p + 1/q = 1$. *Then*

$$|\mu(fg)| \leq \|f\|_p \|g\|_q. \tag{2.16}$$

Proof: It is sufficient to prove this when $\|f\|_p = 1$ and $\|g\|_p = 1$. However by the lemma

$$|f(x)||g(x)| \leq \frac{1}{p}|f(x)|^p + \frac{1}{q}|g(x)|^q. \tag{2.17}$$

Integrate.

This lemma shows that if g is in $L^q(\mu)$, with $1 < q < \infty$, then the linear functional defined by $f \mapsto \mu(fg)$ is continuous on $L^p(\mu)$, where $1 < p < \infty$ with $1/p + 1/q = 1$. This shows that each element of $L^q(\mu)$ defines an element of the dual space of $L^p(\mu)$. It may be shown that every element of the dual space arises in this way. Thus the dual space of $L^p(\mu)$ is $L^q(\mu)$, for $1 < p < \infty$.

Notice that we also have a Hölder inequality in the limiting case:

$$|\mu(fg)| \leq \|f\|_1 \|g\|_\infty. \tag{2.18}$$

This shows that every element g of $L^\infty(\mu)$ defines an element of the dual space of $L^1(\mu)$. It may be shown that if μ is σ-finite, then $L^\infty(\mu)$ is the dual space of $L^1(\mu)$.

On the other hand, each element f of $L^1(\mu)$ defines an element of the dual space of $L^\infty(\mu)$. However in general this does not give all elements of the dual space of $L^\infty(\mu)$.

The most important spaces are L^1, L^2, and L^∞. The nicest by far is L^2, since it is a Hilbert space. The space L^1 is also common, since it measures the total amount of something. The space L^∞ goes together rather naturally with L^1. Unfortunately, the theory of the spaces L^1 and L^∞ is more delicate than the theory of the spaces L^p with $1 < p < \infty$. Ultimately this is because the spaces L^p with $1 < p < \infty$ have better convexity properties.

Here is a brief summary of the facts about duality. The dual space of a Banach space is the space of continuous linear scalar functions on the Banach space. The dual space of a Banach space is a Banach space. Let $1/p + 1/q = 1$, with $1 \leq p < \infty$ and $1 < q \leq \infty$. (Require that μ be σ-finite when $p = 1$.) Then the dual of the space $L^p(X, \mu)$ is the space $L^q(X, \mu)$. The dual of $L^\infty(X, \mu)$ is not in general equal to $L^1(X, \mu)$. Typically $L^1(X, \mu)$ is not the dual space of anything.

The fact that is often used instead is that the dual of $C_0(X)$ is $M(X)$. However in general the space $C_0(X)$ is considerably smaller than $L^\infty(X, \mu)$.

Correspondingly, the space $M(X)$ (which consists of signed Radon measures of finite variation) is quite a bit larger than $L^1(X, \mu)$. That is, for each g in $L^1(X, \mu)$ the functional $f \mapsto \mu(fg)$ is continuous on $C_0(X)$ and may be identified with a signed measure $\sigma(f) = \mu(fg)$. But there are many other elements σ of the dual space of $C_0(X)$ that are not given by functions in this way. For example, if $X = \mathbf{R}$, μ is Lebesgue measure, and $\sigma(f) = f(0)$ is point mass at the origin, then σ is not given by integration with a function.

2.4 Orlicz spaces

It is helpful to place the theory of L^p spaces in a general context. Clearly, the theory depends heavily on the use of the functions x^p for $p \geq 1$. This is a convex function. The generalization is to use a more or less arbitrary convex function.

Let $H(x)$ be a continuous function defined for all $x \geq 0$ such that $H(0) = 0$ and such that $H'(x) > 0$ for $x > 0$. Then H is an increasing function. Suppose that $H(x)$ increases to infinity as x increases to infinity. Finally, suppose that $H''(x) \geq 0$. This implies convexity.

Example: $H(x) = x^p$ for $p > 1$.

Example: $H(x) = e^x - 1$.

Example: $H(x) = (x + 1)\log(x + 1)$.

Define the size of f by $\mu(H(|f|))$. This is a natural notion, but it does not have good scaling properties. So we replace f by f/c and see if we can make the size of this equal to one. The c that accomplishes this will be the norm of f.

This leads to the official definition of the *Orlicz norm*

$$\|f\|_H = \inf\{c > 0 \mid \mu\left(H\left(|f/c|\right)\right) \leq 1. \tag{2.19}$$

It is not difficult to show that if this norm is finite, then we can find a c such that

$$\mu\left(H\left(|f/c|\right)\right) = 1. \tag{2.20}$$

Then the definition takes the simple form

$$\|f\|_H = c, \tag{2.21}$$

where c is defined by the previous equation.

It is not too difficult to show that this norm defines a Banach space $L_H(\mu)$. The key point is that the convexity of H makes the norm satisfy the triangle inequality.

Theorem 2.7 *The Orlicz norm satisfies the triangle inequality*

$$\|f + g\|_H \leq \|f\|_H + \|g\|_H. \tag{2.22}$$

Proof: Let $c = \|f\|_H$ and $d = \|g\|_H$. Then by the fact that H is increasing and convex

$$H\left(\left|\frac{f+g}{c+d}\right|\right) \le H\left(\frac{c}{c+d}\left|\frac{f}{c}\right| + \frac{d}{c+d}\left|\frac{g}{d}\right|\right) \le \frac{c}{c+d}H\left(\left|\frac{f}{c}\right|\right) + \frac{d}{c+d}H\left(\left|\frac{g}{d}\right|\right). \tag{2.23}$$

Integrate. This gives

$$\mu\left(H\left(\left|\frac{f+g}{c+d}\right|\right)\right) \le 1. \tag{2.24}$$

Thus $\|f+g\|_H \le c+d$.

Notice that this result is a generalization of Minkowski's inequality. So we see that the idea behind L^p spaces is convexity. The convexity is best for $1 < p < \infty$, since then the function x^p has second derivative $p(p-1)x^{p-2} > 0$. (For $p = 1$ the function x is still convex, but the second derivative is zero, so it not strictly convex.)

One can also try to create a duality theory for Orlicz spaces. For this it is convenient to make the additional assumptions that $H'(0) = 0$ and $H''(x) > 0$ and $H'(x)$ increases to infinity.

The dual function to $H(x)$ is a function $K(y)$ called the *Legendre transform*. The definition of $K(y)$ is

$$K(y) = xy - H(x), \tag{2.25}$$

where x is defined implicitly in terms of y by $y = H'(x)$.

This definition is somewhat mysterious until one computes that $K'(y) = x$. Then the secret is revealed: The functions H' and K' are inverse to each other. Furthermore, the Legendre transform of $K(y)$ is $H(x)$.

Example: Let $H(x) = x^p/p$. Then $K(y) = y^q/q$, where $1/p + 1/q = 1$.

Example: Let $H(x) = e^x - 1 - x$. Then $K(y) = (y+1)\log(y+1) - y$.

Lemma 2.3 *Let $H(x)$ have Legendre transform $K(y)$. Then for all $x \ge 0$ and $y \ge 0$*

$$xy \le H(x) + K(y). \tag{2.26}$$

Proof: Fix y and consider the function $xy - H(x)$. Since $H'(x)$ is increasing to infinity, the function rises and then dips below zero. It has its maximum where the derivative is equal to zero, that is, where $y - H'(x) = 0$. However by the definition of Legendre transform, the value of $xy - H(x)$ at this point is $K(y)$.

Theorem 2.8 *(Hölder's inequality) Suppose that H and K are Legendre transforms of each other. Then*

$$|\mu(fg)| \le 2\|f\|_H \|g\|_K. \tag{2.27}$$

Proof: It is sufficient to prove this when $\|f\|_H = 1$ and $\|g\|_K = 1$. However by the lemma

$$|f(x)||g(x)| \le H(|f(x)|) + K(|g(y)|). \tag{2.28}$$

Integrate.

This is just the usual derivation of Hölder's inequality. However if we take $H(x) = x^p/p$, $K(y) = y^q/q$, then the H and K norms are not quite the usual L^p and L^q, but instead multiples of them. This explains the extra factor of 2. In any case we see that the natural context for Hölder's inequality is the Legendre transform for convex functions. For more on this general subject, see Appendix H (Young-Orlicz spaces) in R. M. Dudley, Uniform Central Limit Theorems, Cambridge University Press, 1999.

2.5 Problems

1. Define the Fourier transform for f in L^1 by defining for each k

$$\hat{f}(k) = \int e^{-ikx} f(x)\, d^n x. \tag{2.29}$$

The inverse Fourier transform is

$$f(x) = \int e^{ikx} \hat{f}(k)\, \frac{d^n k}{(2\pi)^n}. \tag{2.30}$$

Show that if f is in L^1, then the Fourier transform is in L^∞ and is continuous. Prove an analogous result for the inverse Fourier transform. Hint: Use the dominated convergence theorem.

2. Define the Fourier transform for f in L^2 by

$$\hat{f}(k) = \lim_{m\to\infty} \int_{|x|\le m} e^{-ikx} f(x)\, d^n x. \tag{2.31}$$

The limit is taken in the L^2 sense. The inverse Fourier transform is defined for $\hat{f}$ in L^2 by

$$f(x) = \lim_{m\to\infty} \int_{|k|\le m} e^{ikx} \hat{f}(k)\, \frac{d^n k}{(2\pi)^n}. \tag{2.32}$$

The Plancherel formula is

$$\int |f(x)|^2\, dx = \int |\hat{f}(k)|^2\, \frac{d^n k}{(2\pi)^n}. \tag{2.33}$$

The Fourier transform and the inverse Fourier transform each send L^2 to L^2. They are isomorphisms (unitary transformations) between complex Hilbert spaces. Give examples of functions in L^2 that are not in L^1 for which these transforms exist, and calculate them explicitly.

3. For certain f in L^2 one can define the Hilbert space derivative $\partial f(x)/\partial x_j$ in L^2 by insisting that $ik_j \hat{f}(k)$ is in L^2 and taking the inverse Fourier transform. Show (in the one dimensional case) that a Hilbert space derivative can exist even when there are points at which the pointwise derivative does not exist. Hint: Try a triangle function.

4. Similarly, one can define the Laplace operator Δ by requiring that the Fourier transform of Δf be $-k^2 \hat{f}(k)$. Show that if $a > 0$, then the equation $(a^2 - \Delta)f = g$ with g in L^2 always has a unique solution f in L^2.

5. In numerical analysis of PDE one wants to solve the backward difference scheme

$$u_{j+1} = u_j + h\Delta u_{j+1}. \tag{2.34}$$

Fix u_0. Take $h = t/j$ and let $j \to \infty$. Prove that u_j converges in the L^2 sense as $j \to \infty$. Hint: Use the Fourier transform and the dominated convergence theorem.

6. Show that the corresponding convergence with the forward difference scheme

$$u_{j+1} = u_j + h\Delta u_j \tag{2.35}$$

has big problems.

7. Fix $a > 0$. For each $m \geq 0$, define the Sobolev space H^m to consist of all functions f in L^2 such that $(a^2 - \Delta)^{\frac{m}{2}} f$ is also in L^2. Show that if $m > n/2$, then each function f in H^m is bounded and continuous. Hint: Show that $\hat{f}$ is in L^1.

8. Let g be in L^2. Take $n < 4$. Show that the f in L^2 that solves $(a^2 - \Delta)f = g$ is bounded and continuous.

9. Show that the condition $n < 4$ in the preceding problem is necessary to get the conclusion. Hint: Look at a function of the radius.

10. Let K be the Legendre transform of H. Show that if $H''(x) > 0$, then also $K''(y) > 0$. What is the relation between these two functions?

Chapter 3

Probability

3.1 Random variables and expectation

Probability has its own terminology. There is a set Ω called the *sample space*. Each element ω of Ω is called an *outcome*. There is also a σ-algebra of measurable functions on Ω denoted by $\mathcal{F}$. A function X in $\mathcal{F}$ is called a *random variable*. Thus it is a number that depends on the outcome of the experiment. For each outcome ω in Ω there is a corresponding experimental number $X(\omega)$.

There is also an integral E that is required to satisfy the additional property

$$E[c] = c. \tag{3.1}$$

That is, the integral of every constant random variable is the same constant. This integral is called the *expectation*, or sometime the *mean*.

Theorem 3.1 *(Jensen's inequality) If X is a random variable and ϕ is a smooth convex function, then*

$$\phi(E[X]) \leq E[\phi(X)]. \tag{3.2}$$

Proof. If ϕ is a convex function, then

$$\phi(a) + \phi'(a)(x - a) \leq \phi(x). \tag{3.3}$$

In particular

$$\phi(E[X]) + \phi'(E[X])(X - E[X]) \leq \phi(X). \tag{3.4}$$

This gives a corresponding inequality for expectations.

Note: Jensen's inequality is true without the hypothesis of smoothness; however this proof exhibits its absolutely elementary nature.

Corollary 3.1 *(inequality of geometric and arithmetic mean) If $Y > 0$, then*

$$\exp(E[\log Y]) \leq E[Y]. \tag{3.5}$$

Corollary 3.2 *(inequality of harmonic and arithmetic mean) If $Y > 0$, then*

$$\frac{1}{E[\frac{1}{Y}]} \leq E[Y]. \tag{3.6}$$

An random variable X with finite L^p norm $E[|X|^p]^{\frac{1}{p}} < \infty$ is said to have a pth moment.

Corollary 3.3 *If $1 \leq p \leq p' < \infty$ and if X has finite $L^{p'}$ norm, then X has finite L^p norm, and*

$$E[|X|^p]^{\frac{1}{p}} \leq E[|X|^{p'}]^{\frac{1}{p'}}. \tag{3.7}$$

The space L^2 is of particular importance. In probability it is common to use the *centered* random variable $X - E[X]$. This is the random variable that measures deviations from the expected value. There is a special terminology in this case. The *variance* of X is

$$\mathrm{Var}(X) = E[(X - E[X])^2]. \tag{3.8}$$

Note the important identity

$$\mathrm{Var}(X) = E[X^2] - E[X]^2. \tag{3.9}$$

There is a special notation that is in standard use. The mean of X is written

$$\mu_X = E[X]. \tag{3.10}$$

The Greek mu reminds us that this is a mean. The variance of X is written

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{3.11}$$

The square root of the variance is the *standard deviation* of X. This is

$$\sigma_X = \sqrt{E[(X - \mu_X)^2]}. \tag{3.12}$$

The Greek sigma reminds us that this is a standard deviation. If we center the random variable and divided by its standard deviation, we get the *standardized* random variable

$$Z = \frac{X - \mu_X}{\sigma_X}. \tag{3.13}$$

The *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \tag{3.14}$$

Note the important identity

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]. \tag{3.15}$$

From the Schwarz inequality we have the following important theorem.

Theorem 3.2

$$|\operatorname{Cov}(X,Y)| \le \sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}. \tag{3.16}$$

Sometimes this is stated in terms of the *correlation coefficient*

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} \tag{3.17}$$

which is the covariance of the standardized random variables. The result is the following.

Corollary 3.4

$$|\rho(X,Y)| \le 1. \tag{3.18}$$

Perhaps the most important theorem in probability is the following. It is a trivial consequence of linearity, but it is the key to the law of large numbers.

Theorem 3.3

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X,Y) + \operatorname{Var}(Y). \tag{3.19}$$

Random variables X and Y are said to be *uncorrelated* if $\operatorname{Cov}(X,Y) = 0$. Note that this is equivalent to the identity $E[XY] = E[X]E[Y]$.

Corollary 3.5 *If X and Y are uncorrelated, then the variances add:*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y). \tag{3.20}$$

Random variables X, Y are said to be *independent* if for every pair of functions f, g for which the expectations exist

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]. \tag{3.21}$$

The following result is so important that it must be stated as a theorem.

Theorem 3.4 *If X, Y are independent L^2 random variables, then they are uncorrelated.*

The notion of independence may be extended to several random variables. Thus, for instance, X, Y, Z are independent if for all suitable f, g, h we have

$$E[f(X)g(Y)h(Z)] = E[f(X)]E[g(Y)]E[h(Z)]. \tag{3.22}$$

3.2 Probability models

Here are some important examples.

Example. Let $S = \{1, 2, \ldots m\}$ be a finite set with m elements. Let $p_1, p_2, \ldots, p_m$ be numbers with $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_m = 1$. We can think of S as a set of outcomes corresponding to a single trial of an experiment with m possible outcomes. Let Y be the random variable defined by $Y(i) = i$. Then for each function f the expectation of the random variable $f(Y)$ is

$$E[f(Y)] = \sum_{i=1}^{m} f(i))p_i. \tag{3.23}$$

This is a sum over m points.

If $m = 6$ and each $p_i = 1/6$, then this can be a model for a throw of a die. Thus Y would be the number shown on the upper face of the die. If $m = 2$ and $p_1 = 1 - p$, $p_2 = p$, then this is a model for a toss of a biased coin with probability p of heads. Then $X = Y - 1$ would score 1 for a head and 0 for a tail.

Example. Let $S = \{1, 2, \ldots m\}$ be a finite set with m elements. Let $p_1, p_2, \ldots, p_m$ be numbers with $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_m = 1$. We can think of S as a set of outcomes corresponding to a single trial of an experiment with m possible outcomes. Let Ω be the set of all finite sequences $\omega_1, \ldots, \omega_N$ of elements of S. This is a finite set of m^N outcomes. We think of Ω as a set of outcomes corresponding to N independent trials of an experiment with m possible outcomes. For each $i = 1, \ldots, N$, let Y_i be the random variable on Ω whose value on ω is ω_i. Consider a random variable $f(Y_1, \ldots, Y_N)$ that depends on $Y_1, \ldots, Y_N$. Then

$$E[f(Y_1, \ldots, Y_N)] = \sum_{i_1=1}^{m} \cdots \sum_{i_N=1}^{m} f(i_1, \ldots, i_N)p_{i_1} \cdots p_{i_N}. \tag{3.24}$$

This is a sum over m^N points. When $m = 6$ and each $p_i = 1/6$ we can think of N independent throws of a die. Thus Y_j would be the number shown on the upper face on the jth throw. When $m = 2$ and $p_1 = 1 - p$, $p_2 = p$, then this is a model for N independent tosses of a biased coin. Then $X_j = Y_j - 1$ would be the indicator of the event that the jth toss results in a head. In any case, the random variables $y_1, Y_2, \ldots, Y_N$ are independent.

Consider some k with $1 \leq k \leq N$. Consider a random variable $f(Y_1, \ldots, Y_k)$ that depends on $Y_1, \ldots, Y_k$. Then it is easy to see that

$$E[f(Y_1, \ldots, Y_k)] = \sum_{i_1=1}^{m} \cdots \sum_{i_k=1}^{m} f(i_1, \ldots, i_k)p_{i_1} \cdots p_{i_k}. \tag{3.25}$$

This is a sum over m^k points. If k is considerably less than N, then the use of this formula can be a considerable economy.

Example. Let $S = \{1, 2, \ldots m\}$ be a finite set with m elements. Let $p_1, p_2, \ldots, p_m$ be numbers with $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_m = 1$. We can think of S as a set of outcomes corresponding to a single trial of an experiment with m possible outcomes. Let Ω be the set of all infinite sequences $\omega_1, \ldots, \omega_k, \ldots$ of elements of S. This is an uncountably infinite set. We think of Ω as a set of outcomes corresponding to infinitely many independent trials of an experiment with m possible outcomes. For each $i = 1, 2, 3 \ldots$, let Y_i be the random variable on S whose value on ω is ω_i. Consider some k with $1 \leq k < \infty$. Consider a random variable $f(Y_1, \ldots, Y_k)$ that depends on $Y_1, \ldots, Y_k$. Then we want the formula

$$E[f(Y_1, \ldots, Y_k)] = \sum_{i_1=1}^{m} \cdots \sum_{i_k=1}^{m} f(i_1, \ldots, i_k) p_{i_1} \cdots p_{i_k}. \tag{3.26}$$

This is a sum over m^k points. However there are also random variables that depend on all the $Y_1, Y_2, Y_3, \ldots$. For such random variables the expectation cannot be calculated by such a simple formula. However it may be shown that there is a unique expectation defined for arbitrary random variables that has the property that for random variables that depend on only finitely many trials the expectation is given by this formula. In this model the random variables $Y_1, Y_2, \ldots, Y_k, \ldots$ are independent.

3.3 The sample mean

In statistics the *sample mean* is used to estimate the population mean.

Theorem 3.5 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{3.27}$$

be their sample mean. Then the expectation of $\bar{X}_n$ is

$$E[\bar{X}_n] = \mu. \tag{3.28}$$

Proof: The expectation of the sample mean $\bar{X}_n$ is

$$E[\bar{X}_n] = E[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n} E[\sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu. \tag{3.29}$$

Theorem 3.6 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{3.30}$$

be their sample mean. Then the standard deviation of $\bar{X}_n$ is

$$\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{3.31}$$

Proof: The variance of the sample mean $\bar{X}_n$ is

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2. \tag{3.32}$$

We can think of these two results as a form of the weak law of large numbers. The law of large numbers is the "law of averages" that says that averaging uncorrelated random variable gives a result that is approximately constant. In this case the sample mean has expectation μ and standard deviation $\sigma/\sqrt{n}$. Thus if n is large enough, it is a random variable with expectation μ and with little variability.

The factor $1/\sqrt{n}$ is both the blessing and the curse of statistics. It is a wonderful fact, since it says that averaging reduces variability. The problem, of course, is that while $1/\sqrt{n}$ goes to zero as n gets larger, it does so rather slowly. So one must somehow obtain a quite large sample in order to ensure rather moderate variability.

The reason the law is called the weak law is that it gives a statement about a fixed large sample size n. There is another law called the strong law that gives a corresponding statement about what happens for all sample sizes n that are sufficiently large. Since in statistics one usually has a sample of a fixed size n and only looks at the sample mean for this n, it is the more elementary weak law that is relevant to most statistical situations.

3.4 The sample variance

The sample mean

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.33}$$

is a random variable that may be used to estimate an unknown population mean μ. In the same way, the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1} \tag{3.34}$$

may be used to estimate an unknown population variance σ^2.

The $n-1$ in the denominator seems strange. However it is due to the fact that while there are n observations X_i, their deviations from the sample mean $X_i - \bar{X}_n$ sum to zero, so there are only $n-1$ quantities that can vary independently. The following theorem shows how this choice of denominator makes the calculation of the expectation give a simple answer.

Theorem 3.7 *Let $X_1, X_2, X_3, \ldots X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let s^2 be the sample variance. Then the expectation of s^2 is*

$$E[s^2] = \sigma^2. \tag{3.35}$$

Proof: Compute

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X}_n+\bar{X}_n-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X}_n)^2+\sum_{i=1}^{n}(\bar{X}_n-\mu)^2. \quad (3.36)$$

Notice that the cross terms sum to zero. Take expectations. This gives

$$n\sigma^2 = E[\sum_{i=1}^{n}(X_i-\bar{X}_n)^2)] + n\frac{\sigma^2}{n}. \quad (3.37)$$

The result then follows from elementary algebra.

The above proof shows how the fact that the sum of the $X_i - \bar{X}_n$ is zero is responsible for the $n-1$ factor. It might seem that $E[s^2] = \sigma^2$ is a conclusive argument for using the $n-1$ in the definition of the sample variance s^2. On the other hand, Jensen's inequality shows that for the sample standard deviation s it is typical that $E[s] < \sigma$. So perhaps it is more a question of elegance than necessity.

3.5 Events and probability

Probability is a special case of expectation! If A is a subset of Ω such that the indicator function 1_A is measurable, then A is called an *event*. Thus when the experiment is conducted, the event A happens or does not happen according to whether the outcome ω is in A or not. The *probability* of an event A is given by the expectation of the indicator indicator random variable:

$$P[A] = E[1_A]. \quad (3.38)$$

This relation between event and random variables is very useful. The value $1_A(\omega) = 1$ if the outcome ω belongs to A, and the value $1_A(\omega) = 0$ if the outcome ω does not belong to A. Thus one scores 1 if the outcome belongs to A, and one scores zero if the outcome does not belong to A.

Events A, B are said to be *independent* if their indicator functions are independent random variables. This is equivalent to saying that $P[A \cap B] = P[A]P[B]$. Similarly, events A, B, C are said to be independent if their indicator functions are independent. This is equivalent to saying that $P[A \cap B] = P[A]P[B]$, $P[B \cap C] = P[B]P[C]$, $P[A \cap C] = P[A]P[C]$, and $P[A \cap B \cap B] = P[A]P[B]P[C]$. There is a similar definition for more than three events.

In realistic probability problems the set theory language is often replaced by an equivalent terminology. An event is thought of as a property of outcomes. The notions of union, intersection, complement are replaced by the logical notations or, and, not. Thus additivity says that for *exclusive* events, the probability that one *or* the other occurs is the *sum* of the probabilities. Also, the probability that an event occurs plus the probability that an event does not occurs is one. The definition of *independent* events is that the probability that one event *and* the other event occur is the *product* of the probabilities.

In probability theory when a property holds almost everywhere the probabilistic terminology is to say that it holds *almost surely*.

If X is a random variable, and S is a Borel set of real numbers, then the event $X \in S$ that X is in S consists of all the outcomes ω such that $X(\omega)$ is in S. Let 1_S be the indicator function defined on the set of real numbers that is 1 for a number in S and 0 for a number not in S. Then the indicator function of the event that X is in S is the random variable $1_S(X)$.

3.6 The sample proportion

In statistics the *sample proportion* f_n is used to estimate the population proportion p.

Theorem 3.8 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Let $N_n = 1_{A_1} + \cdots + 1_{A_n}$ be the number of events that occur. Let*

$$f_n = \frac{N_n}{n} \tag{3.39}$$

be the sample frequency. Then the expectation of f_n is

$$E[f_n] = p. \tag{3.40}$$

Theorem 3.9 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Assume that each pair A_i, A_j of events $i \neq j$ are independent. Let $N_n = 1_{A_1} + \cdots + 1_{A_n}$ be the number of events that occur. Let*

$$f_n = \frac{N_n}{n} \tag{3.41}$$

be the sample frequency. Then the standard deviation of f_n is

$$\sigma_{f_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}. \tag{3.42}$$

Proof: The variance of 1_{A_1} is $p - p^2 = p(1-p)$.

In order to use this theorem, the following remark upper bound on the standard deviation for one observation is fundamental.

$$\sqrt{p(1-p)} \leq \frac{1}{2}. \tag{3.43}$$

This means that one can figure out an upper bound on the standard deviation of the sample frequency without knowing the population probability. Often, because of the central limit theorem, one wants to think of a reasonable bound on the probable error to be 2 standard deviations. Thus the memorable form of this result is

$$2\sigma_{f_n} \leq \frac{1}{\sqrt{n}}. \tag{3.44}$$

This is not quite as important, but for many practical problems there is also a useful lower bound for the standard deviation of one observation. If $1/10 \leq p \leq 9/10$, then $3/10 \leq \sqrt{p(1-p)}$.

3.7 Orthogonal projection

One of the most useful concepts in Hilbert space theory is that of orthogonal projection.

Theorem 3.10 *Let $\mathcal{H}$ be a real Hilbert space. Let $\mathcal{M}$ be a closed subspace of $\mathcal{H}$. Let x be a vector in $\mathcal{H}$. Then there exists a unique vector w such that*

1. *w is in $\mathcal{M}$, and*
2. *$x - w$ is orthogonal to $\mathcal{M}$.*

The condition that $x - w$ is orthogonal to $\mathcal{M}$ says that the inner product $\langle (x - w), z\rangle = 0$ for all z in $\mathcal{M}$. This in turn says that $\langle w, z\rangle = \langle x, z\rangle$ for all z in $\mathcal{M}$. If we set $w = Px$, the two conditions may be stated:

1. Px is in $\mathcal{M}$, and
2. For every z in $\mathcal{M}$
$$\langle Px, z\rangle = \langle x, z\rangle. \tag{3.45}$$

Theorem 3.11 *Let $\mathcal{H}$ be a real Hilbert space. Let $\mathcal{M}$ be a closed subspace of $\mathcal{H}$. Let x be a vector in $\mathcal{H}$. Then the orthogonal projection w of x on $\mathcal{M}$ is the vector in $\mathcal{M}$ closest to x, that is, w is in $\mathcal{M}$ and*

$$\|x - w\|^2 \leq \|x - u\|^2 \tag{3.46}$$

for all u in $\mathcal{M}$.

Proof: Let F be defined on $\mathcal{M}$ by

$$F(u) = \|x - u\|^2. \tag{3.47}$$

Let c be the infimum of F over the set $\mathcal{M}$. Let u_n be a sequence of vectors in $\mathcal{M}$ such that $F(u_n) \to c$. We have the identity

$$F(\frac{u_m + u_n}{2}) + \frac{1}{4}\|u_n - u_m\|^2 = \frac{1}{2}F(u_m) + \frac{1}{2}F(u_n). \tag{3.48}$$

This is a very strong form of convexity. Since $\mathcal{M}$ is convex, this implies that

$$c + \frac{1}{4}\|u_n - u_m\|^2 \leq \frac{1}{2}F(u_m) + \frac{1}{2}F(u_n). \tag{3.49}$$

From this it follows that u_n is a Cauchy sequence. Therefore there exists w in $\mathcal{M}$ such that $u_n \to w$. Furthermore, $F(u_n) \to F(w)$, so $F(w) = c$.

Fix z in $\mathcal{M}$. The function $F(w + tz)$ has its minimum at $t = 0$. Since

$$F(w + tz) = F(w) - 2t\langle x - w, z\rangle + t^2\|z\|^2, \tag{3.50}$$

the derivative at $t = 0$ is $-2\langle x - w, z\rangle = 0$.

3.8 Conditional expectation and probability

We begin with an expectation E defined on the positive elements of a σ-algebra $\mathcal{F}$ of random variables on Ω. As usual, it extends to random variables such that either the positive or negative part has finite expectation. For the moment we mainly deal with L^1 random variables such that $E[|X|] < \infty$, so that both the positive and negative part have finite expectation. We shall also deal with the smaller class of L^2 random variables such that $E[X^2] < \infty$, that is, have finite variance.

Let $\mathcal{F}_1$ be a σ-algebra of functions contained in $\mathcal{F}$. We wish to define the *conditional expectation* of X given $\mathcal{F}_1$. We do this first in the special case $E[X^2] < \infty$, that is, when X is in L^2.

In this case there is a particularly simple *Hilbert space definition* of conditional expectation. If X is in L^2, then $E[X \mid \mathcal{F}_1]$ is defined to be the orthogonal projection of X onto the closed subspace $\mathcal{F}_1 \cap L^2$. This translates to the conditions

1. $E[X \mid \mathcal{F}_1]$ is in $\mathcal{F}_1 \cap L^2$, and

2. For every Z in $\mathcal{F}_1 \cap L^2$

$$E[E[X \mid \mathcal{F}_1]Z] = E[XZ] \tag{3.51}$$

Another way of thinking of the conditional expectation is in terms of best approximation. If X is in L^2, then the conditional expectation $E[X \mid \mathcal{F}_1]$ is the function in the σ-algebra of functions $\mathcal{F}_1$ that best approximates X in the L^2 sense.

The intuitive meaning of conditional expectation is the following. Recall that the expectation $E[X]$ is a mathematical prediction made before the experiment is conducted about averages involving X. Suppose now that information is available about the values of the random variables in $\mathcal{F}_1$. Then $E[X \mid \mathcal{F}_1]$ is a mathematical prediction that makes use of this new partial knowledge. It is itself random, since it depends on the values of the random variables in $\mathcal{F}_1$.

Sometimes, instead of indicating the σ-algebra of random variables $\mathcal{F}_1$, one just indicates a collection of random variables that generate the σ-algebra. Thus $E[X \mid Y_1, \ldots, Y_k]$ is the conditional expectation given the σ-algebra of functions generated by $Y_1, \ldots, Y_k$. This σ-algebra consists of all random variables $h(Y_1, \ldots, Y_k)$ where h is a Borel function.

Example: Throw a pair of dice. Let X be the number on the first die; let Y be the number on the second die. Then $E[X + Y] = 7$, but

$$E[X + Y \mid X] = X + \frac{7}{2}. \tag{3.52}$$

This is because the fact that X and Y are independent gives

$$E[(X + Y)h(X)] = E[(X + \frac{7}{2})h(X)] \tag{3.53}$$

for every random variable $h(X)$ in the σ-algebra generated by X. This is intuitive. If you know the number on the first die, then the only averaging is over the possible numbers on the second die.

Example: Throw a pair of dice. Let X be the number on the first die; let Y be the number on the second die. Then

$$E[X \mid X + Y] = \frac{X + Y}{2}. \tag{3.54}$$

This is because it is clear from symmetry that

$$E[\frac{X + Y}{2} h(X + Y)] = E[X h(X + Y)] \tag{3.55}$$

for every random variable $h(X+Y)$ in the σ-algebra generated by $X+Y$. This is intuitive. If you know the sum of the numbers on the dice, then you constrain the possibilities for the numbers on the first die.

Lemma 3.1 *If X is in L^2 and $X \geq 0$, then $E[X \mid \mathcal{F}_1] \geq 0$.*

Proof: Suppose $X \geq 0$. Let $W = E[X \mid \mathcal{F}_1]$. Then W is in $\mathcal{F}_1$. Let A be the event that $W < 0$. Then 1_A is in $\mathcal{F}_1$. Thus $0 \leq E[W 1_A]$. This implies that $W \geq 0$.

Lemma 3.2 *If X is in L^2, then*

$$E[|E[X \mid \mathcal{F}_1]|] \leq E[|X|]. \tag{3.56}$$

Thus the conditional expectation is continuous as a map from L^1 to L^1.

Proof: This follows from applying the previous lemma to the positive and negative parts of X. Thus if $X = X_+ - X_-$ and $|X| = X_+ + X_-$ and $X_+ \geq 0$, $X_- \geq 0$, then

$$E[X \mid \mathcal{F}_1] = E[X_+ \mid \mathcal{F}_1] - E[X_- \mid \mathcal{F}_1], \tag{3.57}$$

so

$$|E[X \mid \mathcal{F}_1]| \leq E[X_+ \mid \mathcal{F}_1] + E[X_- \mid \mathcal{F}_1] = E[|X| \mid \mathcal{F}_1]. \tag{3.58}$$

This lemma allows us to extend the definition of conditional expectation from $L^1 \cap L^2$ to all of L^1. It is then easy to prove the following characterization.

Let $\mathcal{F}_1$ be a σ-algebra of random variables contained in $\mathcal{F}$. Let X be a random variable such that $E[|X|] < \infty$, that is, X is in L^1. The *conditional expectation* of X given $\mathcal{F}_1$ is a random variable denoted by $E[X \mid \mathcal{F}_1]$. It is characterized by the two requirements

1. $E[X \mid \mathcal{F}_1]$ is in $\mathcal{F}_1$, and

2. For every bounded random variable Z in $\mathcal{F}_1$

$$E[E[X \mid \mathcal{F}_1] Z] = E[XZ]. \tag{3.59}$$

There are many other useful properties of conditional expectation. The following theorem gives one such property.

Theorem 3.12 *If X is in L^1 and if W is a bounded random variable in $\mathcal{F}_1$, then*

$$E[WX \mid \mathcal{F}_1] = WE[X \mid \mathcal{F}_1]. \tag{3.60}$$

Thus random variables in $\mathcal{F}_1$ act like constants when one takes conditional expectations with respect to $\mathcal{F}_1$.

Once we have conditional expectation, then we also have conditional probability. We define $P[A \mid \mathcal{F}_1] = E[1_A \mid \mathcal{F}_1]$. This is a random variable with the additional property that $0 \leq P[A \mid \mathcal{F}_1] \leq 1$. It is a revised estimate of the probability of A when the additional information about the values of the random variables in $\mathcal{F}_1$ is given.

3.9 Problems

1. Consider the experiment of throwing a die n times. The results are $X_1, \ldots, X_n$. Then $E[f(X_i)] = \frac{1}{6}\sum_{k=1}^{6} f(k)$, and the X_i are independent. Find the mean μ and standard deviation σ of each X_i.

2. Consider the dice experiment. Take $n = 25$. Find the mean $\mu_{\bar{X}}$ of the sample mean $\bar{X}$. Find the standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$.

3. Perform the dice experiment with $n = 25$ and get an outcome ω. Record the 25 numbers. Report the sample mean $\bar{X}(\omega)$. Report the sample standard deviation $s(\omega)$.

4. Consider the experiment of throwing a die n times. Let A_i be the event that the ith throw gives a number in the range from 1 to 4. Find the mean μ and standard deviation σ of the indicator function of each A_i.

5. Consider the same dice experiment. Take $n = 50$. Find the mean μ_f of the sample proportion f. Find the standard deviation σ_f of the sample proportion f.

6. Perform the dice experiment with $n = 50$ and get an outcome ω. Record the 50 events A_i. Report the sample proportion $f(\omega)$.

7. Consider a random sample of size n from a very large population. The question is to find what proportion p of people in the population have a certain opinion. The proportion in the sample who have the opinion is f. How large must n be so that the standard deviation of f is guaranteed to be no larger than one percent?

8. Let X_1 and X_2 be the numbers resulting from throwing two dice. Let A be the event that X_1 is odd, let B be the event that X_2 is odd, and let C

be the event that $X_1 + X_2$ is odd. Show that A, B are independent, A, C are independent, and B, C are independent. Show that A, B, C are not independent.

9. Consider independent random variables $X_1, \ldots, X_n, \ldots$. For notational convenience, consider the centered random variables $Y_i = X_i - \mu$, so that $E[Y_i] = 0$. Let $\sigma^2 = E[Y_i^2]$ and $q^4 = E[Y_i^4]$. Prove that

$$E[\bar{Y}_k^4] = \frac{1}{k^4}[kq^4 + 3k(k-1)\sigma^4]. \tag{3.61}$$

10. In the proceeding problem, show that

$$E[\sum_{k=n}^{\infty} \bar{Y}_k^4] \leq \frac{1}{2}q^4 \frac{1}{(n-1)^2} + 3\sigma^4 \frac{1}{(n-1)}. \tag{3.62}$$

In terms of the original X_i this says that there is a constant C such that

$$E[\sum_{k=n}^{\infty} (\bar{X}_k - \mu)^4] \leq C\frac{1}{n}. \tag{3.63}$$

Thus if n is large, then all the sample means $\bar{X}_k$ for $n \leq k < \infty$ are likely to be close to μ, in some average sense. This is a form of the strong law of large numbers. Compare with the weak law

$$E[(\bar{X}_n - \mu)^2] = \sigma^2 \frac{1}{n}, \tag{3.64}$$

which only shows that, for each fixed n, the sample mean $\bar{X}_n$ is very likely to be close to μ.

11. In the preceding problem, show that

$$\lim_{k \to \infty} \bar{X}_k = \mu \tag{3.65}$$

almost surely. This is the usual elegant statement of the strong law of large numbers.

12. Consider independent identically distributed random variables $X_1, \ldots, X_n, \ldots$ with finite variance. We know from the weak law of large numbers that

$$E[(\bar{X}_n - \mu)^2] = \sigma^2 \frac{1}{n}, \tag{3.66}$$

Use this to prove that

$$P[|\bar{X}_n - \mu| \geq t] \leq \frac{\sigma^2}{nt^2}. \tag{3.67}$$

Thus for large n the probability that the sample mean $\bar{X}_n$ deviates from the population mean μ by t or more is small. This is another form of the weak law of large numbers.

13. Consider independent random variables $X_1, \ldots, X_n, \ldots$ with finite fourth moments. We have seen that

$$E[\sum_{j=n}^{\infty} (\bar{X}_j - \mu)^4] \leq \frac{C}{n}. \tag{3.68}$$

Show that

$$P[\max_{n \leq j < \infty} |\bar{X}_j - \mu| \geq t] \leq \frac{C}{nt^4}. \tag{3.69}$$

Thus for large n the probability that there exists some j with $n \leq j < \infty$ such that the sample mean $\bar{X}_j$ deviates from the population mean μ by t or more is small. This is another form of the strong law of large numbers.

14. Let X be a random variable. Let B_j be an indexed family of disjoint measurable sets whose union is Ω. Consider the σ-algebra of measurable functions generated by the indicator functions 1_{B_j}. Find the conditional expectation of X with respect to this algebra.

15. Find $P[A \mid 1_B]$.

16. Find $P[A \mid 1_B, 1_C]$.

17. Suppose that X, Y have a joint density $\rho(x, y)$, so that

$$E[f(X, Y)] = \int \int f(x, y)\rho(x, y)\, dx\, dy. \tag{3.70}$$

Show that

$$E[f(X, Y) \mid Y] = \frac{\int f(x, Y)\rho(x, Y)\, dx}{\int \rho(x, Y)\, dx}. \tag{3.71}$$

Chapter 4

Random walk and martingales

4.1 Symmetric simple random walk

Let $X_0 = x$ and

$$X_{n+1} = X_n + \xi_{n+1}. \tag{4.1}$$

The ξ_i are independent, identically distributed random variables such that $P[\xi_i = \pm 1] = 1/2$. The probabilities for this random walk also depend on x, and we shall denote them by P_x. We can think of this as a fair gambling game, where at each stage one either wins or loses a fixed amount.

Let T_y be the first time $n \geq 1$ when $X_n = y$. Let $\rho_{xy} = P_x[T_y < \infty]$ be the probability that the walk starting at x ever gets to y at some future time.

First we show that $\rho_{12} = 1$. This says that in the fair game one is almost sure to eventually get ahead by one unit. This follows from the following three equations.

The first equation says that in the first step the walk either goes from 1 to 2 directly, or it goes from 1 to 0 and then must go from 0 to 2. Thus

$$\rho_{12} = \frac{1}{2} + \frac{1}{2}\rho_{02}. \tag{4.2}$$

The second equation says that to go from 0 to 2, the walk has to go from 0 to 1 and then from 1 to 2. Furthermore, these two events are independent. Thus

$$\rho_{02} = \rho_{01}\rho_{12}. \tag{4.3}$$

The third equation says that $\rho_{01} = \rho_{12}$. This is obvious.

Thus $\rho = \rho_{12}$ satisfies

$$\rho = \frac{1}{2}\rho^2 + \frac{1}{2}. \tag{4.4}$$

This is a quadratic equation that can also be written as $\rho^2 - 2\rho + 1 = 0$, or $(\rho - 1)^2 = 0$. Its only solution is $\rho = 1$.

How long does it take, on the average, to get ahead by this amount? Let $m_{xy} = E_x[T_y]$, the expected time that the random walk takes to get to y, starting at x. We have just seen that if $x = 1$, then $T_2 < \infty$ with probability one. Let us do the same kind of computation for $m_{12} = E_1[T_2]$.

The first equation says that in the first step the walk either goes from 1 to 2 directly, or it goes from 1 to 0 and then must go from 0 to 2. Thus

$$m_{12} = \frac{1}{2}(1 + m_{02}) + \frac{1}{2}1. \tag{4.5}$$

The second equation says that to go from 0 to 2, the walk has to go from 0 to 1 and then from 1 to 2. Thus

$$m_{02} = m_{01} + m_{12}. \tag{4.6}$$

The third equation says that $m_{01} = m_{12}$. This is obvious.

Thus $m = m_{12}$ satisfies

$$m = \frac{1}{2}(1 + 2m) + \frac{1}{2}. \tag{4.7}$$

This is a linear equation that can also be written as $m = 1+m$. Its only solution is $m = \infty$. Thus we have seen that $m_{01} = \infty$.

At first this seems strange, but it is quite natural. In the fair game there is a small probability of a bad losing streak. It takes a long time to recover from the losing streak and eventually get ahead. Infinitely long, on the average.

The calculation above uses a rather subtle property of random walk. Namely, it was assumed that after the walk accomplishes the task of going from 0 to 1, then it has an equally difficult task of going from 1 to 2. Now this is delicate, because the walk going from 1 to 2 starts out at a random time, not at a fixed time. Namely, if we start the walk at 0 and set $T = T_1$ as the first time the walk gets to 1, then from that time on the walk is $X_T, X_{T+1}, X_{T+2}, \ldots$ with $X_T = 1$ by definition. The point is that this has the same distribution as the walk $X_0, X_1, X_2, \ldots$ with $X_0 = 1$. This fact is called the *strong Markov property*.

The strong Markov property has the following statement. Start the walk out at x and let $T = T_y$. Let B be a set of random walk paths. We must prove that

$$P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B \mid T < \infty] = P_y[(X_0, X_1, X_2, \ldots) \in B]. \tag{4.8}$$

This says that the walk does not care when and how it got to y for the first time; from that point on it moves just as if it were started there at time zero.

The proof is the following. We write

$$P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B, T < \infty] = \sum_{n=1}^{\infty} P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B, T = n]. \tag{4.9}$$

However

$$P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B, T = n] = P_x[(X_n, X_{n+1}, X_{n+2} \ldots) \in B, T = n]. \tag{4.10}$$

Now the event $T = n$ depends only on the walk $X_1, X_2, \ldots, X_n$, which depends only on the $\xi_1, \xi_2, \ldots, \xi_n$. On the other hand, the walk $X_n, X_{n+1}, X_{n+2}, \ldots$ with $X_n = y$ depends only on the $\xi_{n+1}, \xi_{n+1}, \ldots$. Thus the events are independent, and we have

$$P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B, T = n] = P_x[(X_n, X_{n+1}, X_{n+2} \ldots) \in B]P[T = n], \tag{4.11}$$

where $X_n = y$. Finally, this is

$$P_x[(X_T, X_{T+1}, X_{T+2}, \ldots) \in B, T = n] = P_y[(X_0, X_1, X_2 \ldots) \in B]P[T = n]. \tag{4.12}$$

The proof is finished by summing over n.

The picture of symmetric simple random walk that emerges is the following. Starting from any point, the probability of eventually getting to any other point is one. However the expected time to accomplish this is infinite. Thus an imbalance in one direction is always compensated, but this random process is incredibly inefficient and can take a huge amount of time to do it.

4.2 Simple random walk

Let $X_0 = x$ and

$$X_{n+1} = X_n + \xi_{n+1}. \tag{4.13}$$

The ξ_i are independent, identically distributed random variables such that $P[\xi_i = 1] = p$, $P[\xi_i = -1] = q$, and $P[\xi = 0] = r$. Here $p + q + r = 1$, so the walk can change position by only one step at a time. The probabilities for this random walk depend on x, and we shall denote them by P_x. We can think of this as a gambling game, where at each stage one either wins or loses a fixed amount.

Let T_y be the first time $n \geq 1$ when $X_n = y$. Let $\rho_{xy} = P_x[T_y < \infty]$ be the probability that the walk starting at x ever gets to y at some future time.

First we calculate ρ_{12}. The calculation again uses three equations.

The first equation says that in the first step the walk either goes from 1 to 2 directly, or it stays at 1 and then must go to 2, or it goes from 1 to 0 and then must go from 0 to 2. Thus

$$\rho_{12} = q\rho_{02} + r\rho_{12} + p. \tag{4.14}$$

The second equation says that to go from 0 to 2, the walk has to go from 0 to 1 and then from 1 to 2. Furthermore, these two events are independent. Thus

$$\rho_{02} = \rho_{01}\rho_{12}. \tag{4.15}$$

The third equation says that $\rho_{01} = \rho_{12}$. This is obvious.

Thus $\rho = \rho_{12}$ satisfies

$$\rho = q\rho^2 + r\rho + p. \tag{4.16}$$

This is a quadratic equation that can also be written as $q\rho^2 + (r-1)\rho + p = 0$, or $(\rho - 1)(q\rho - p) = 0$. If $q > 0$ its solutions are $\rho = 1$ and $\rho = p/q$.

If $q \le p$, then it is clear that the probability $\rho_{01} = 1$, except in the case $p = q = 0$. The game is even or favorable, so one is eventually ahead.

If $p < q$, then it would seem plausible that the probability is $\rho_{01} = p/q$. The game is unfavorable, and there is some chance that there is an initial losing streak from which one nevers recovers. There are a number of ways of seeing that this is the correct root. If the root were one, then the process starting at any number $x \le 0$ would be sure to eventually reach 1. However according to the strong law of large numbers the walk $X_n/n \to p - q < 0$ as $n \to \infty$. Thus the walk eventually becomes negative and stays negative. This is a contradiction.

How long does it take, on the average, to get ahead by this amount? Let $m_{xy} = E_x[T_y]$, the expected time that the random walk takes to get to y, starting at x. We have just seen that if $x = 1$ and $p \ge q$ (not both zero), then $T_2 < \infty$ with probability one. Let us do the same kind of computation for $m_{12} = E_1[T_2]$.

The first equation says that in the first step the walk either goes from 1 to 2 directly, or it remains at 1 and then goes to 2, or it goes from 1 to 0 and then must go from 0 to 2. Thus

$$m_{12} = q(1 + m_{02}) + r(1 + m_{12}) + p1. \tag{4.17}$$

The second equation says that to go from 0 to 2, the walk has to go from 0 to 1 and then from 1 to 2. Thus

$$m_{02} = m_{01} + m_{12}. \tag{4.18}$$

Here we are implicitly using the strong Markov property.

The third equation says that $m_{01} = m_{12}$. This is obvious.

Thus $m = m_{12}$ satisfies

$$m = q(1 + 2m) + r(1 + m) + p. \tag{4.19}$$

This is a linear equation that can also be written as $m = 1 + (1 + q - p)m$. Its solutions are $m = \infty$ and $m = 1/(p - q)$.

If $p \le q$, then it is clear that the expectation $m_{12} = \infty$. The game is even or unfavorable, so it can take a very long while to get ahead, if ever.

If $q < p$, then it would seem plausible that the expected number of steps to get ahead by one is $m_{12} = 1/(p - q)$. The game is favorable, so one should have win relatively soon, on the average. There are a number of ways of seeing that this is the correct solution. One way would be to use the identity

$$E_1[T_2] = \sum_{k=0}^{\infty} k P_1[T_2 = k] = \sum_{k=1}^{\infty} P_1[T_2 \ge k]. \tag{4.20}$$

Then one can check that if $q < p$, then the probabilities $P_1[T_2 \ge k]$ go rapidly to zero, so the series converges. This rules out an infinite expectation.

4.3 Gambler's ruin: fair game

Let $X_0 = x$ and

$$X_{n+1} = X_n + \xi_{n+1}. \tag{4.21}$$

The ξ_i are independent, identically distributed random variables such that $P[\xi_i = 1] = p$, $P[\xi_i = -1] = q$, and $P[\xi_i = 0] = r$. Here $p + q + r = 1$, so the walk can change position by only one step at a time. The probabilities for this random walk depend on x, and we shall denote them by P_x.

We can think of this as a gambling game, where at each stage one either wins or loses a fixed amount. However now we want to consider the more realistic case when one wants to win a certain amount, but there is a cap on the possible loss. Thus take $a < x < b$. The initial capital is x. We want to play until the time T_b when the earnings achieve the desired level b or until the time T_a when we are broke. Let T be the minimum of these two times. Thus T is the time the game stops.

There are a number of ways to solve the problem. One of the most elegant is the martingale method. A martingale is a fair game.

We first want to see when the game is fair. In this case the conditional expectation of the winnings at the next stage are the present fortune. This conditional expectation is

$$E[X_{n+1} \mid X_n] = E[X_n + \xi_{n+1} \mid X_n] = X_n + p - q. \tag{4.22}$$

So this just says that $p = q$.

It follows from the equation $E[X_{n+1} \mid X_n] = X_n]$ that

$$E[X_{n+1}] = E[X_n]. \tag{4.23}$$

Therefore, starting at x we have

$$E_x[X_n] = x. \tag{4.24}$$

The expected fortune stays constant.

Let $T \wedge n$ be the minimum of T and n. Then $X_{T \wedge n}$ is the game stopped at T. It is easy to show that this too is a fair game. Either one has not yet won or lost, and the calculation of the conditional expectation is as before, or one has won or lost, and future gains and losses are zero. Therefore

$$E_x[X_{T \wedge n}] = x. \tag{4.25}$$

Now the possible values of the stopped game are bounded: $a \leq X_{T \wedge n} \leq b$ for all n. Furthermore, since $T < \infty$ with probability one, the limit as $n \to \infty$ of $X_{T \wedge n}$ is X_T. It follows from the dominated convergence theorem (described below) that

$$E_x[X_T] = x. \tag{4.26}$$

Thus the game remains fair even in the limit of infinite time.

From this we see that

$$aP_x[X_T = a] + bP_x[X_T = b] = x. \tag{4.27}$$

It is easy then to calculate, for instance, that the probability of winning is

$$P_x[X_T = b] = \frac{x-a}{b-a}. \tag{4.28}$$

4.4 Gambler's ruin: unfair game

If the original game is not fair, then we want to modify it to make it fair. The new game will be that game whose accumulated winnings at stage n is $f(X_n)$. The conditional expectation of the winnings in the new game at the next stage are the present fortune of this game. Thus

$$E[f(X_{n+1}) \mid X_n] = f(X_n). \tag{4.29}$$

This says that $qf(X_n - 1) + rf(X_n) + pf(X_n + 1) = f(X_n)$. One solution of this is to take f to be a constant function. However this will not give a very interesting game. Another solution is obtained by trying an exponential $f(z) = a^z$. this gives $q/a + r + pa = 1$. If we take $a = q/p$ this gives a solution. Thus in the following we take

$$f(X_n) = \left(\frac{q}{p}\right)^{X_n} \tag{4.30}$$

as the fair game. If $p < q$, so the original game is unfair, then in the new martingale game rewards being ahead greatly and penalized being behind only a little.

It follows from this equation that

$$E[f(X_{n+1})] = E[f(X_n)]. \tag{4.31}$$

Therefore, starting at x we have

$$E_x[f(X_n)] = f(x). \tag{4.32}$$

The expected fortune in the new martingale game stays constant.

Notice that to play this game, one has to arrange that the gain or loss at stage $n + 1$ is

$$f(X_{n+1}) - f(X_n) = \left(\left(\frac{q}{p}\right)^{\xi_{n+1}} - 1\right) f(X_n). \tag{4.33}$$

If, for instance, $p < q$, then the possible gain of $(q-p)/p\, f(X_n)$ is greater than the possible loss of $(q-p)/q\, f(X_n)$; the multiplicative factor of q/p makes this a fair game.

Let $T \wedge n$ be the minimum of T and n. Then $X_{T\wedge n}$ is the game stopped at T. It is easy to show that this too is a fair game. Therefore the same calculation shows that

$$E_x[f(X_{T\wedge n})] = f(x). \tag{4.34}$$

Now if $p < q$ and $f(x) = (q/p)^x$, we have that $a \le x \le b$ implies $f(a) \le f(x) \le f(b)$. It follows that $f(a) \le f(X_{T\wedge n}) \le f(b)$ is bounded for all n with bounds that do not depend on n. This justifies the passage to the limit as $n \to \infty$. This gives the result

$$E_x[f(X_T)] = f(x). \tag{4.35}$$

The game is fair in the limit.

From this we see that

$$\left(\frac{q}{p}\right)^a P_x[X_T = a] + \left(\frac{q}{p}\right)^b P_x[X_T = b] = \left(\frac{q}{p}\right)^x. \tag{4.36}$$

It is easy then to calculate, for instance, that the probability of winning is

$$P_x[X_T = b] = \frac{(q/p)^x - (q/p)^a}{(q/p)^b - (q/p)^a}. \tag{4.37}$$

If we take $p < q$, so that we have a losing game, then we can recover our previous result for the probability of eventually getting from x to b in the random walk by taking $a = -\infty$. Then we get $(q/p)^{x-b} = (p/q)^{b-x}$.

4.5 Martingales

The general definition of a martingale is the following. We have a sequence of random variables $\xi_1, \xi_2, \xi_3, \ldots$. The random variable Z_n is a function of $\xi_1, \ldots, \xi_n$, so that

$$Z_n = h_n(\xi_1, \ldots, \xi_n). \tag{4.38}$$

The martingale condition is that

$$E[Z_{n+1} \mid \xi_1, \ldots, \xi_n] = Z_n. \tag{4.39}$$

Thus a martingale is a fair game. The expected value at the next stage, given the present and past, is the present fortune.

It follows by a straightforward calculation that

$$E[Z_{n+1}] = E[Z_n]. \tag{4.40}$$

In particular $E[Z_n] = E[Z_0]$.

The fundamental theorem about martingales says that if one applies a gambling scheme to a martingale, then the new process is again a martingale. That is, no gambling scheme can convert a fair game to a game where one has an unfair advantage.

Theorem 4.1 *Let $\xi_1, \xi_2, \xi_3, \ldots$ be a sequence of random variables. Let $X_0, X_1, X_2, \ldots$ be a martingale defined in terms of these random variables, so that X_n is a function of $\xi_1, \ldots, \xi_n$. Let W_n be a gambling scheme, that is, a function of $\xi_1, \ldots, \xi_n$. Let $Z_0, Z_1, Z_2, \ldots$ be a new process such that $Z_{n+1} - Z_n = W_n(X_{n+1} - X_n)$. Thus the gain in the new process is given by the gain in the original process modified by the gambling scheme. Then $Z_0, Z_1, Z_2, \ldots$ is also a martingale.*

Proof: The condition that X_n is a martingale is that the expected gain $E[X_{n+1} - X_n \mid \xi_1, \ldots, \xi_n] = 0$. Let

$$W_n = g_n(\xi_1, \ldots, \xi_n) \tag{4.41}$$

be the gambling scheme. Then

$$E[Z_{n+1} - Z_n \mid \xi_1, \ldots, \xi_n] = E[W_n(X_{n+1} - X_n) \mid \xi_1, \ldots, \xi_n]. \tag{4.42}$$

On the other hand, since W_n is a function of $\xi_1, \ldots, \xi_n$, this is equal to

$$W_n E[X_{n+1} - X_n \mid \xi_1, \ldots, \xi_n] = 0. \tag{4.43}$$

Example: Say that $X_n = \xi_1 + \xi_2 + \cdots \xi_n$ is symmetric simple random walk. Let the gambling scheme $W_n = 2^n$. That is, at each stage one doubles the amount of the bet. The resulting martingale is $Z_n = \xi_1 + 2\xi_2 + 4\xi_3 + \cdots + 2^{n-1}\xi_n$. This game gets wilder and wilder, but it is always fair. Can one use this double the bet game to make money? See below.

One important gambling scheme is to quit gambling once some goal is achieved. Consider a martingale $X_0, X_1, X_2, \ldots$. Let T be a time with the property that the event $T \leq n$ is defined by $\xi_1, \ldots, \xi_n$. Such a time is called a *stopping time*. Let the gambling scheme W_n be 1 if $n < T$ and 0 if $T \leq n$. Then the stopped martingale is given by taking $Z_0 = X_0$ and $Z_{n+1} - Z_n = W_n(X_{n+1} - X_n)$. Thus if $T \leq n$, $Z_{n+1} - Z_n = 0$, while if $n < T$ then $Z_{n+1} - Z_n = X_{n+1} - X_n$. As a consequence, if $T \leq n$, then $Z_n = X_T$, while if $n < T$, then $Z_n = X_n$. This may be summarized by saying that $Z_n = X_{T \wedge n}$, where $T \wedge n$ is the minimum of T and n. In words: the process no longer changes after time T.

Corollary 4.1 *Let T be a stopping time. Then if X_n is a martingale, then so is the stopped martingale $Z_n = X_{n \wedge T}$.*

It might be, for instance, that T is the first time that the martingale X_n belongs to some set. Such a time is a stopping time. Then the process $X_{T \wedge n}$ is also a martingale.

Example: A gambler wants to use the double the bet martingale $Z_n = \xi_1 + 2\xi_2 + 4\xi_3 + \cdots + 2^{n-1}\xi_n$ to get rich. The strategy is to stop when ahead. The process $Z_{T \wedge n}$ is also a martingale. This process will eventually win one unit. However, unfortunately for the gambler, at any particular non-random time n the stopped process is a fair game. $Z_{T \wedge n}$ is either 1 with probability

$1 - 1/2^n$ or $1 - 2^n$ with probability $1/2^n$. It looks like an easy win, most of the time. But a loss is a disaster.

If Z_n is a martingale, then the processes $Z_{T \wedge n}$ where one stops at the stopping time T is also a martingale. However what if $T < \infty$ with probability one, there is no limit on the time of play, and one plays until the stopping time? Does the game remain fair in the limit of infinite time?

Theorem 4.2 *If $T < \infty$ and if there is a random variable $Y \geq 0$ with $E[Y] < \infty$ such that for all n we have $|Z_{T \wedge n}| \leq Y$, then $E[Z_{T \wedge n}] \to E[Z_T]$ as $n \to \infty$.*

This theorem, of course, is just a corollary of the dominated convergence theorem. In the most important special case the dominating function Y is just a constant. This says that for a bounded martingale we can always pass to the limit. In general the result can be false, as we shall see in the following examples.

Example. Let X_n be symmetric simple random walk starting at zero. Then X_n is a martingale. Let T be the first n with $X_n = b > 0$. The process $X_{T \wedge n}$ that stops when b is reached is a martingale. The stopped process is a martingale. However the infinite time limit is not fair! In fact $X_T = b$ by definition. A fair game is converted into a favorable game. However this is only possible because the game is unbounded below.

Example. Let X_n be simple random walk starting at zero, not symmetric. Then $(q/p)^{X_n}$ is a martingale. Let $b > 0$ and T be the first $n \geq 1$ with $X_n = b$. Then $(q/p)^{X_{T \wedge n}}$ is a martingale. If $p < q$, an unfavorable game, then this martingale is bounded. Thus we can pass to the limit. This gives $(q/p)^x = (q/p)^b P[T_b < \infty]$. On the other hand, if $q < p$, then the martingale is badly unbounded. It is not legitimate to pass to the limit. And in fact $(q/p)^x \neq (q/p)^b P[T_b < \infty] = (q/p)^b$.

Example: Consider again the double the bet gambling game where one quits when ahead by one unit. Here $Z_{T \wedge n} = 1$ with probability $1 - 1/2^m$ and $Z_{T \wedge n} = 1 - 2^n$ with probability $1/2^n$. Eventually the gambler will win, so the $Z_T = 1$. This limiting game is no longer a martingale, it is favorable to the gambler. However one can win only by having unlimited credit. This is unrealistic in this kind of game, since the losses along the way can be so huge.

4.6 Problems

1. For the simple random walk starting at zero, find the mean and standard deviation of X_n.

2. For the simple random walk starting at 1, compute ρ_{11}, the probability of a return to the starting point.

3. For the simple random walk starting at 1, compute m_{11}, the expected time to return to the starting point.

4. A gambler plays roulette with the same stake on each play. This is simple random walk with $p = 9/19$ and $q = 10/19$. The player has deep pockets, but must win one unit. What is the probability that he does this in the first three plays? What is the probability that he ever does this?

5. A gambler plays roulette with the same stake on each play. This is simple random walk with $p = 9/19$ and $q = 10/19$. The player strongly desires to win one unit. But he will be thrown out of the game if he is ever behind by 1000 units. What is the the probability that he wins one unit? What is his expected gain?

6. For the simple random walk with $p+q+r = 1$ and $p \neq q$ show that $X_n - (p-q)n$ is a martingale. This is the game in which an angel compensates the player for the average losses. Then let $a < x < b$ and let T be the first time the walk starting at x gets to a or b. Show that the stopped martingale $X_{T\wedge n} - (p-q)T \wedge n$ is a martingale. Finally, use $E_x[T] < \infty$ to show that $E_x[X_T] = (p-q)E_x[T] + x$. Compute $E_x[T]$ explicitly.

7. For the symmetric simple random walk with $p + q + r = 1$ and $p = q$ show that $X_n^2 - (1-r)n$ is a martingale. The average growth of X_n^2 is compensated to make a fair game. Then let $a < x < b$ and let T be the first time the symmetric walk starting at x gets to a or b. Show that the stopped martingale $X_{T\wedge n}^2 - (1-r)T \wedge n$ is a martingale. Finally, use $E_x[T] < \infty$ to show that for the symmetric walk $E_x[X_T^2] = (1-r)E_x[T] + x^2$. Compute $E_x[T]$ explicitly.

Chapter 5

The integral: construction

5.1 The Daniell construction

A vector lattice L is a set of real functions defined on a set X. It must satisfy the following properties:

1. L is a vector space of functions.

2. L is a lattice of functions.

It is not required that the vector lattice contain the constant functions. It is also not required that the vector lattice be closed under pointwise limits.

Example: Consider the interval $X = (0, 1]$ of real numbers. Consider functions that have the property that for some $k = 1, 2, 3, \ldots$ the function has the form

$$f(x) = \sum_{j=0}^{2^k-1} c_j 1_{(j/2^k,(j+1)/2^k]}(x). \tag{5.1}$$

These are step functions based on binary intervals. They form a vector lattice L. In this example it does contain the constant functions.

An *elementary integral* is a real function μ defined on a vector lattice L. It must satisfy the following properties:

1. μ is linear.

2. μ sends positive functions to positive real numbers.

3. μ is continuous under pointwise monotone convergence.

Notice that the values of μ on the elements of L are always finite. The condition of pointwise monotone convergence may be taken to say that if h_n is in L and if $h_n \downarrow 0$ pointwise, then $\mu(h_n) \to 0$. This immediately implies certain other forms of monotone convergence. For instance, if each f_n is in L and if f

is in L and if $f_n \uparrow f$ pointwise, then $\mu(f_n) \to \mu(f)$. The second form follows from the first form by taking $h_n = f - f_n$.

Example: Continue the above example. Define the integral of such a function by

$$\mu(f) = \sum_{j=0}^{2^k-1} c_j \frac{1}{2^k}. \tag{5.2}$$

This is the corresponding Riemann sum. It satisfies the monotone convergence, but this is not completely obvious. Thus it is an elementary integral.

Theorem 5.1 *Let L be a vector lattice and let μ be an elementary integral. Then μ has an extension to a larger vector lattice $\bar{L}^1$ (called the vector lattice of summable functions) on which μ continues to satisfy all the properties of an elementary integral. Furthermore, μ satisfies a form of the monotone convergence theorem on $\bar{L}^1$: If the f_n are summable functions with $\mu(f_n) \leq M < \infty$ and if $f_n \uparrow f$, then f is summable with $\mu(f_n) \to \mu(f)$.*

This theorem is called the Daniell construction of the integral. It may be found in such texts as L. H. Loomis, An Introduction to Abstract Harmonic Analysis, van Nostrand, New York, 1953 and H. L. Royden, Real Analysis, third edition, Macmillan, New York, 1988.

Here is a sketch of a proof, following the treatment in Loomis. It is very abbreviated, but it may help exhibit the overall plan. Then one can look at Loomis or another reference for details.

Begin with a vector lattice L and an elementary integral μ. Let $L\uparrow$ be the set of all pointwise limits of increasing sequences of elements of L. These functions are allowed to take on the value $+\infty$. Similarly, let $L\downarrow$ be the set of all pointwise limits of decreasing sequences of L. These functions are allowed to take on the value $-\infty$. Note that the functions in $L\downarrow$ are the negatives of the functions in $L\uparrow$.

For h in $L\uparrow$, take $h_n \uparrow h$ with h_n in L and define $\mu(h) = \lim_n \mu(h_n)$. The limit of the integral exists because this is a monotone sequence of numbers. Similarly, if g in $L\downarrow$, take $g_n \downarrow g$ with g_n in L and define $\mu(g) = \lim_n \mu(g_n)$. It may be shown that these definitions are independent of the particular sequences of functions in L that are chosen.

It is not hard to show that upward monotone convergence is true for functions in $L\uparrow$. Similarly, downward monotone convergence is true for functions in $L\downarrow$.

Here is the argument for upward monotone convergence. Say that the h_n are in $L\uparrow$ and $h_n \to h$ as $n \to \infty$. For each n, let g_{nm} be a sequence of functions in L such that $g_{nm} \uparrow h_n$ as $m \to \infty$. Let $u_n = g_{1n} \vee g_{2n} \vee \cdots \vee g_{nn}$. Then u_n is in L and we have the squeeze inequality

$$g_{in} \leq u_n \leq h_n \tag{5.3}$$

for $1 \leq i \leq n$. As $n \to \infty$ the $g_{in} \uparrow h_i$ and the $h_n \uparrow h$. Furthermore, as $i \to \infty$ the $h_i \uparrow h$. By the squeeze inequality $u_n \uparrow h$. From the squeeze inequality we

get

$$\mu(g_{in}) \le \mu(u_n) \le \mu(h_n) \tag{5.4}$$

for $1 \le i \le n$. By definition of the integral on $L\uparrow$ we can take $n \to \infty$ and get $\mu(h_i) \le \mu(h) \le \lim_n \mu(h_n)$. Then we can take $i \to \infty$ and get $\lim_i \mu(h_i) \le \mu(h) \le \lim_n \mu(h_n)$. This shows that the integrals converge to the correct value.

A function f is said to be in $\bar{L}^1$ (or to be summable) if for every $\epsilon > 0$ there is a function g in $L\downarrow$ and a function h in $L\uparrow$ such that $g \le f \le h$, $\mu(g)$ and $\mu(f)$ are finite, and $\mu(h) - \mu(g) < \epsilon$. Then it is not hard to show that there is a number $\mu(f)$ that is the supremum of all the $\mu(g)$ for g in $L\downarrow$ with $g \le f$ and that is also the infimum of all the $\mu(h)$ for h in $L\uparrow$ with $f \le h$.

It is not hard to show that the set $\bar{L}^1$ of summable functions is a vector lattice and that μ is a positive linear functional on it. The crucial point is that there is also a monotone convergence theorem. This theorem says that if the f_n are summable functions with $\mu(f_n) \le M < \infty$ and if $f_n \uparrow f$, then f is summable with $\mu(f_n) \to \mu(f)$.

Here is the proof. We may suppose that $f_0 = 0$. Since the summable functions $\bar{L}^1$ are a vector space, each $f_n - f_{n-1}$ for $n \ge 1$ is summable. Consider $\epsilon > 0$. Choose h_n in $L\uparrow$ for $n \ge 1$ such that $f_n - f_{n-1} \le h_n$ and such that

$$\mu(h_n) \le \mu(f_n - f_{n-1}) + \epsilon/2^n. \tag{5.5}$$

Let $s_n = \sum_{i=1}^n h_i$ in $L\uparrow$. Then $f_n \le s_n$ and

$$\mu(s_n) \le \mu(f_n) + \epsilon \le M + \epsilon. \tag{5.6}$$

Also $s_n \uparrow s$ in $L\uparrow$ and $f \le s$. and so by monotone convergence for $L\uparrow$

$$\mu(s) \le \lim_n \mu(f_n) + \epsilon \le M + \epsilon. \tag{5.7}$$

Now pick m so large that $f_m \le f$ satisfies $\mu(s) < \mu(f_m) + \frac{3}{2}\epsilon$. Then pick r in $L\downarrow$ with $r \le f_m$ so that $\mu(f_m) \le \mu(r) + \frac{1}{2}\epsilon$. Then $r \le f \le s$ with $\mu(s) - \mu(r) < 2\epsilon$. Since ϵ is arbitrary, this proves that f is summable. Since $f_n \le f$, it is clear that $\lim_n \mu(f_n) \le \mu(f)$. On the other hand, the argument has shown that for each $\epsilon > 0$ we can find s in $L\uparrow$ with $f \le s$ and $\mu(f) \le \mu(s) \le \lim_n \mu(f_n) + \epsilon$. Since ϵ is arbitrary, we conclude that $\mu(f) \le \lim_n \mu(f_n)$.

Remarks on the proof:

1. The construction of the integral is a two stage process. One first extends the integral from L to the increasing limits in $L\uparrow$ and to the decreasing limits in $L\downarrow$. Then to define the integral on $\bar{L}^1$ one approximates from above by functions in $L\uparrow$ or from below by functions in $L\downarrow$.

2. The proof of the monotone convergence theorem for the functions in $L\uparrow$ and for the functions in $L\downarrow$ is routine. However the proof of the monotone convergence theorem for the functions in $\bar{L}^1$ is much deeper. In particular, it uses in a critical way the fact that the sequence of functions is indexed by a countable set of n. Thus the errors in the approximations can be estimated by $\epsilon/2^n$, and these sum to the finite value ϵ.

Example: Continue the example. The extended integral is the Lebesgue integral defined on the space $\bar{L}^1$ of summable Lebesgue measurable functions on the interval $(0, 1]$. Let us show that the integral of the indicator function of a countable set Q is 0. This will involve a two-stage process. Let q_j, $j = 1, 2, 3, \ldots$ be an enumeration of the points in Q. Fix $\epsilon > 0$. For each j, find a binary interval B_j of length less than $\epsilon/2^j$ such that q_j is in the interval. The indicator function 1_{B_j} of each such interval is in L. Let $h = \sum_j 1_{B_j}$. Then h is in $L\uparrow$ and $\mu(h) \leq \epsilon$. Furthermore, $0 \leq 1_Q \leq h$. This is the first stage of the approximation. Now consider a sequence of $\epsilon > 0$ values that approach zero, and construct in the same way a sequence of h_ϵ such that $0 \leq 1_Q \leq h_\epsilon$ and $\mu(h_\epsilon) \leq \epsilon$. This is the second stage of the approximation. This shows that the integral of 1_Q is zero.

Notice that this could not have been done in one stage. There is no way to cover Q by finitely many binary intervals of small total length. It was necessary first find infinitely many binary intervals that cover Q and have small total length, and only then let this length approach zero.

In the following corollary we consider a vector lattice L. Let $L\uparrow$ consist of pointwise limits of increasing limits from L, and let $L\downarrow$ consist of pointwise limits of decreasing sequences from L. Similarly, let $L\uparrow\downarrow$ consist of pointwise limits of decreasing sequences from $L\uparrow$, and let $L\downarrow\uparrow$ consist of pointwise limits of increasing sequences from $L\downarrow$.

Corollary 5.1 *Let L be a vector lattice and let μ be an elementary integral. Consider its extension μ to $\bar{L}^1$. Then for every f in $\bar{L}^1$ there is a g in $L\downarrow\uparrow$ and an h in $L\uparrow\downarrow$ with $g \leq f \leq h$ and $\mu(f - g) = 0$ and $\mu(h - g) = 0$.*

This corollary says that if we identify functions in $\bar{L}^1$ when the integral of the absolute value of the difference is zero, then all the functions that we ever will need may be taken, for instance, from $L\uparrow\downarrow$. Of course this class is not closed under pointwise limits. So the natural domain of definition of the integral is perhaps that in the following theorem.

Theorem 5.2 *Let L be a vector lattice and let μ be an elementary integral. Let $\bar{L}^1$ be the summable functions. Let $\mathcal{F}_0$ be the smallest class of functions that contains L and is closed under pointwise monotone limits. Then $\mathcal{F}_0$ is a vector lattice. Let L^1 be the intersection of $\bar{L}^1$ with $\mathcal{F}_0$. If f is in $\mathcal{F}_0$ and $0 \leq f \leq g$ and g is in L^1, then f is also in L^1, and $0 \leq \mu(f) \leq \mu(g)$. If f is in $\mathcal{F}_0$ and $0 \leq f$ and f is not summable, then define $\mu(f) = \infty$. Then the monotone convergence theorem holds: If each f_n is in $\mathcal{F}_0$ and if $0 \leq f_n \uparrow f$, then $\mu(f_n) \to \mu(f)$.*

The proof of this theorem is not terribly difficult, but the technicalities of the proof are not so important for a first introduction. See Loomis for this proof.

Note: In the context of the above theorem we can define the integral more generally. If f in $\mathcal{F}_0$ has positive part $f_+ = f \vee 0$ and negative part $f_- = -(f \wedge 0)$, we can write $f = f_+ - f_-$. Then $\mu(f) = \mu(f_+) - \mu(f_-)$ is defined except when there is an $\infty - \infty$ problem. It is easy to see that if f is in $\mathcal{F}_0$, then f is summable if and only if $\mu(f)$ is defined and $|\mu(f)| \leq \mu(f_+) + \mu(f_-) = \mu(|f|) < \infty$.

Note: All the functions in $L\uparrow$ and all the functions in $L\downarrow$ are in $\mathcal{F}_0$. So are the functions in $L\uparrow\downarrow$ and the functions in $L\downarrow\uparrow$.

Note: There can be summable functions in $\bar{L}^1$ that are not in $\mathcal{F}_0$ and hence not in L^1. However there are already so many functions in $\mathcal{F}_0$ that we need not consider any others. From now on we shall only consider the integral defined for functions in $\mathcal{F}_0$.

Example: Continue the example. The integral on L^1 is the Lebesgue integral defined for Borel measurable functions on the interval $(0, 1]$. There are so many Borel measurable functions that this integral is sufficient for almost all practical purposes.

We say that an elementary integral is *σ-finite* if there is a sequence of functions $f_n \geq 0$ in L for each n and with $f_n \uparrow 1$. Note that this is a condition on the vector lattice. However, since $\mu(f)$ is finite for each f in L, it says that 1 is a pointwise monotone limit of positive functions with finite elementary integral.

Example: Continue the above example. The elementary integral is not only σ-finite, it is finite. This means that 1 is already in L. For the example $\mu(1) = 1$.

Theorem 5.3 *Let L be a vector lattice and let μ be a σ-finite elementary integral. Let $\mathcal{F}$ be the smallest σ-algebra of measurable functions containing L. Then μ extends to a σ-finite integral defined on $\mathcal{F}$ (except where there is an $\infty - \infty$ problem). Furthermore, the extension is unique.*

Sketch of proof: Consider the space $\mathcal{F}_0$ constructed in the previous theorem. The only problem is that $\mathcal{F}_0$ might not contain the constant functions. However if the vector lattice is σ-finite, then $\mathcal{F}_0 = \mathcal{F}$.

5.2 Product integral and Lebesgue integral

In this section and the next section we shall see that the expectation for infinitely many tosses of a fair coin is essentially the same as the Lebesgue integral for functions on the unit interval. In the first case the elementary integral can be defined for functions that depend on the results of finitely many tosses. In the second case the elementary integral can be defined for binary step functions. In either case, the extension to an integral on a σ-algebra of measurable functions is accomplished by verifying the hypotheses of the last theorem of the previous section.

The path we shall take is to define the integral for coin tossing in this section, and then transfer it to the Lebesgue integral on the unit interval in the following section.

In the following it will be convenient to denote the set of natural numbers by $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$. The set of strictly positive natural numbers is then $\mathbf{N}_+ = \{1, 2, 3, 4 \ldots\}$.

Let $\Omega = \{0, 1\}^{\mathbf{N}_+}$ be the set of all infinite sequences of zeros and ones. We shall be interested in functions f defined on Ω. For each $k = 1, 2, 3, \ldots$ consider the σ-algebra $\mathcal{F}_k$ of functions on Ω that depend only on the first k elements of the sequence, that is, such that $f(\omega) = g(\omega_1, \ldots, \omega_k)$ for some function g on $\mathbf{R}^k$.

This σ-algebra will be finite dimensional with dimension 2^k. Say that for each $k = 1, 2, 3, \ldots$ we have an integral μ_k defined for the functions in this σ-algebra $\mathcal{F}_k$. We shall say that these integrals are consistent if whenever $j \leq k$ and f is in $\mathcal{F}_j \subset \mathcal{F}_k$, then $\mu_j(f) = \mu_k(f)$.

If we have such a consistent family of integrals, then we can define a positive linear functional μ on the space L that is the union of all the $\mathcal{F}_k$ for $k = 1, 2, 3, \ldots$. Thus L consists of all functions f for which there exists some k (depending on the function) such that f depends only on the first k elements of the sequence. The functional is defined by $\mu(f) = \mu_k(f)$ whenever f is in $\mathcal{F}_k$. It is well-defined on account of the consistency.

Example: If the function f is in $\mathcal{F}_k$, let

$$\mu_k(f) = \sum_{\omega(1)=0}^{1} \cdots \sum_{\omega(k)=0}^{1} f(\omega) \frac{1}{2^k}. \tag{5.8}$$

Then this defines a consistent family. This example describes the expectation for independent of tosses of a fair coin.

Theorem 5.4 *Let $\Omega = \{0, 1\}^{\mathbf{N}+}$ be the set of all infinite sequences of zeros and ones. Say that for each k there is an integral μ_k on the space $\mathcal{F}_k$ of functions f on Ω that depend only on the first k elements of the sequence. Say that these integrals have the consistency property. Let L be the union of the $\mathcal{F}_k$ for $k = 1, 2, 3, \ldots$. Thus the μ_k for $k = 1, 2, 3, \ldots$ define a positive linear functional μ on L given by $\mu(f) = \mu_k(f)$ for f in $\mathcal{F}_k$. Let $\mathcal{F}$ be the σ-algebra of functions generated by L. Then μ extends to an integral on $\mathcal{F}$.*

Proof: The space L is a vector lattice that contains the constant functions, and μ is a positive linear functional on L. The only remaining thing to prove is that μ satisfies the monotone convergence theorem on L. Then the general extension theorem applies.

Let f_n be in L and $f_n \downarrow 0$. We must show that $\mu(f_n) \to 0$. If there were a fixed k such that each f_n depended only on the first k coordinates, then this would be trivial. However this need not be the case.

The idea is to prove the converse. Say there were an $\epsilon > 0$ such that $\mu(f_n) \geq \epsilon$. Then we must prove that there is an infinite sequence $\bar{\omega}$ such that $f_n(\bar{\omega})$ does not converge to zero.

Suppose that $f_n \leq M$ for all n. Let $\alpha > 0$ and $\beta > 0$ be such that $\alpha + M\beta < \epsilon$. Consider the event A_n that $f_n \geq \alpha$. The probability of this event satisfies $\mu(A_n) = \mu(f_n \geq \alpha) \geq \beta$. Otherwise, we would have

$$\mu(f_n) \leq \mu(f_n 1_{A_n^c}) + \mu(f_n 1_{A_n}) \leq \alpha + M\beta < \epsilon. \tag{5.9}$$

This would be a contradiction.

This argument shows that the events A_n are a decreasing sequence of events with $\mu(A_n) \geq \beta > 0$. The next step is to use this to show that the intersection of the A_n is not empty. This says that there is a sequence $\bar{\omega}$ that belongs to each A_n.

Let $\bar{\omega}(1), \ldots, \bar{\omega}(k)$ be a sequence of k zeros and ones. Let $\mu[A_n, \omega(1) = \bar{\omega}(1), \ldots, \omega(k) = \bar{\omega}(k)]$ be the probability of A_n together with this initial sequence. We claim that one can choose the sequence such that for each n this probability is bounded below by

$$\mu[A_n, \omega(1) = \bar{\omega}(1), \ldots, \omega(k) = \bar{\omega}(k)] \geq \beta \frac{1}{2^k}. \tag{5.10}$$

The proof is by induction. The statement is true for $k = 0$. Suppose the statement is true for $k - 1$. Note that

$$\begin{aligned} \mu[A_n, \omega(1) = \bar{\omega}(1), \ldots, \omega(k-1) = \bar{\omega}(k-1)] = \\ \mu[A_n, \omega(1) = \bar{\omega}(1), \ldots, \omega(k-1) = \bar{\omega}(k-1), \omega(k) = 1] \\ + \quad \mu[A_n, \omega(1) = \bar{\omega}(1), \ldots, \omega(k-1) = \bar{\omega}(k-1), \omega(k) = 0]. \end{aligned} \tag{5.11}$$

Suppose that there exists an n such that the first probability on the right with $\omega(k) = 1$ is less than $\beta/2^k$. Suppose also that there exists an n such that the second probability on the right with $\omega(k) = 0$ is less than $\beta/2^k$. Then, since the events are decreasing, there exists an n such that both probabilities are simultaneously less than $\beta/2^k$. But then the probability on the left would be less than $\beta/2^{k-1}$ for this n. This is a contradiction. Thus one of the two suppositions must be false. This says that one can choose $\bar{\omega}(k)$ equal to 1 or to 0 in such a way that the desired inequality holds. This completes the inductive proof of the claim.

It follows that the sequence $\bar{\omega}$ is in each A_n. The reason is that for each n there is a k such that A_n depends only on the first k elements of the sequence ω. Since the probability of A_n together with the first k elements of the sequence $\bar{\omega}$ specified is strictly positive, there must exist at least one sequence ω that agrees with $\bar{\omega}$ in the first k places that belongs to A_n. It follows that every sequence ω that agrees with $\bar{\omega}$ in the first k places is in A_n. In particular, $\bar{\omega}$ is in A_n.

The last argument proves that there is a sequence $\bar{\omega}$ that belongs to each A_n. In other words, there is a sequence $\bar{\omega}$ with $f_n(\bar{\omega}) \geq \alpha > 0$. This completes the proof of the theorem.

Remark: One could consider the space Ω_0 of all infinite sequences of zeros and ones, such that each sequence in the space is eventually all zeros or eventually all ones. This is a countable set. It is easy to construct an elementary integral for functions on Ω_0 that describes coin tossing for finite sequences of arbitrary length. However it would not satisfy the monotone convergence theorem. It is instructive to consider why the proof of the above theorem does not work for Ω_0. The theorem is saying something deep about the way one can model infinite processes. In particular, it is essential that Ω is uncountable.

Theorem 5.5 *Let $\Omega = \{0,1\}^{\mathbf{N}_+}$ be the set of all infinite sequences of zeros and ones. Fix p with $0 \leq p \leq 1$. If the function f on Ω is in the space $\mathcal{F}_k$ of functions that depend only on the first k values of the sequence, let*

$$\mu_k(f) = \sum_{\omega(1)=0}^{1} \cdots \sum_{\omega(k)=0}^{1} f(\omega) p^{\omega(1)}(1-p)^{1-\omega(1)} \cdots p^{\omega(k)}(1-p)^{1-\omega(k)}. \tag{5.12}$$

This defines a consistent family μ_k, $k = 1, 2, 3, \ldots$, of integrals. Let $\mathcal{F}$ be the σ-algebra of functions generated by the $\mathcal{F}_k$ for $k = 1, 2, 3, \ldots$. Then there is an integral μ associated with the σ-algebra $\mathcal{F}$ such that μ agrees with μ_k on each $\mathcal{F}_k$.

This theorem describes the expectation for a sequence of independent coin tosses where the probability of heads on each toss is p and the probability of tails on each toss is $1 - p$. The special case $p = 1/2$ describes a fair coin.

5.3 Image integrals

Let $\mathcal{F}$ be a σ-algebra of measurable functions on X. Let $\mathcal{G}$ be a σ-algebra of measurable functions on Y. A function $\phi : X \to Y$ is called a *measurable map* if for every g in $\mathcal{G}$ the composite function $g(\phi)$ is in $\mathcal{F}$.

Given an integral μ defined on $\mathcal{F}$, and given a measurable map $\phi : X \to Y$, there is an integral $\phi[\mu]$ defined on $\mathcal{G}$. It is given by

$$\phi[\mu](g) = \mu(g(\phi)). \tag{5.13}$$

It is called the *image* of the integral μ under ϕ.

This construction is important in probability theory. If X is a random variable, that is, a measurable function from Ω to $\mathbf{R}$, then it may be regarded as a measurable map. The image of the expectation under X is an integral ν_X on the Borel σ-algebra called the *distribution* of X. We have the identity.

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)\, d\nu_X(x). \tag{5.14}$$

Theorem 5.6 *Let $0 \leq p \leq 1$. Define the product expectation for coin tossing on infinite sequences ω from $\mathbf{N}_+$ to $\{0, 1\}$ as in the theorem. Here p is the probability of heads on each single toss. Let*

$$\phi(\omega) = \sum_{k=1}^{\infty} \omega(k) \frac{1}{2^k}. \tag{5.15}$$

Then the image expectation $\phi[\mu]$ is an expectation defined on Borel functions on the unit interval $[0, 1]$.

When $p = 1/2$ (the product expectation for tossing of a fair coin) this expectation is called the *Lebesgue integral*. However note that there are many other integrals, for the other values of p. We have the following amazing fact. For each p there is an integral defined for functions on the unit interval. If $p \neq p'$ are two different parameters, then these is a measurable set that has measure 1 for the p measure and measure 0 for the p' measurable. The set can be defined as the set of coin tosses for which the sample means converge to the number p. This result shows that these integrals each live in a different world.

We denote the Lebesgue integral for functions on the unit interval by

$$\lambda(f) = \int_0^1 f(u)\,du. \tag{5.16}$$

It is then easy to define the Lebesgue integral for functions defined on the real line $\mathbf{R}$ as the image of an integral for functions defined on the unit interval $[0,1]$. For instance, the integral could be defined for $f \geq 0$ by

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 f(\ln(\frac{u}{1-u}))\frac{1}{u(1-u)}\,du. \tag{5.17}$$

This says that the integral given by dx is the image of the integral given by $1/(u(1-u))\,du$ under the transformation $x = \ln(u/(1-u))$. Notice that this transformation has inverse $u = 1/(1+e^{-x})$. It is a transformation that is often used in statistics to relate a problem on the unit interval $(0,1)$ to a problem on the line.

Once we have the Lebesgue integral defined for Borel functions on the line, we can construct a huge family of other integrals, also defined on Borel functions on the line. These are called Lebesgue-Stieltjes integrals.

Theorem 5.7 *Let F be an increasing right continuous function on $\mathbf{R}$. Then there exists a measure μ defined on the Borel σ-algebra $\mathcal{B}$ such that*

$$\mu((a,b]) = F(b) - F(a). \tag{5.18}$$

Proof: Let $m = \inf F$ and let $M = \sup F$. For $m < y < M$ let

$$G(y) = \sup\{x \mid F(x) < y\}. \tag{5.19}$$

We can compare the least upper bound $G(y)$ with an arbitrary upper bound c. Thus $G(y) \leq c$ is equivalent to the condition that for all x, $F(x) < y$ implies $x \leq c$. This in turn is equivalent to the condition that for all x, $c < x$ implies $y \leq F(x)$. Since F is increasing and right continuous, it follows that this in turn is equivalent to the condition that $y \leq F(c)$.

It follows that $a < G(y) \leq b$ is equivalent to $F(a) < y \leq F(b)$. Thus G is a kind of inverse to F.

Let λ be Lebesgue measure on the interval (m, M). Let $\mu = G[\lambda]$ be the image of this Lebesgue measure under G. Then

$$\mu((a,b]) = \lambda((F(a), F(b)]) = F(b) - F(a), \tag{5.20}$$

so μ is the desired measure.

5.4 Problems

1. Consider the space $\Omega = \{0,1\}^{\mathbf{N}_+}$ with the measure μ that describes fair coin tossing. Let N_3 be the random variable that describes the number

of heads in the first three tosses. Draw the graph of the corresponding function on the unit interval. Find the area under the graph, and check that this indeed gives the expectation of the random variable.

2. Imagine a space $\Omega = \{1, 2, \ldots, m\}^{\mathbf{N}_+}$ of all infinite sequences with values in a finite set. Again let $\mathcal{F}_k$ be the space of real functions on Ω that depend only on the first k elements of the sequence. Show that a consistent family of integrals μ_k on $\mathcal{F}_k$ extends to an integral μ associated with the σ-algebra of functions generated by all the $\mathcal{F}_k$.

3. Consider the interval $[0, 1]$ of real numbers. Consider functions of the form

$$f(x) = \sum_{j=0}^{2^k-1} c_j 1_{(j/2^k,(j+1)/2^k]}(x) \tag{5.21}$$

for some k. This is a step function. Define the integral of such a function by

$$\lambda(f) = \sum_{j=0}^{2^k-1} c_j \frac{1}{2^k}. \tag{5.22}$$

This is the corresponding Riemann sum. Show directly that this integral satisfies the monotone convergence theorem and thus may be extended to an integral

$$\lambda(f) = \int_0^1 f(u)\, du \tag{5.23}$$

defined for Borel functions on the unit interval. Hint: This is a transcription of the theorem on the construction of the coin tossing expectation for $p = 1/2$. Carry out the proof in the new framework. The problem is to construct a real number given by a binary expansion. There is a new difficulty: It is desirable that this expansion not have all zeros from some point on. Perhaps one should make it easier to choose ones. Choose a one every time at least $1/3$ of the previous probability bound is associated with the right hand interval. Does this help?

4. Let

$$\mu(g) = \int_{-\infty}^{\infty} g(t)w(t)\, dt \tag{5.24}$$

be a integral defined by a density $w(t)$ with respect to Lebesgue measure dt. Let $\phi(t)$ be a suitable smooth function that is increasing or decreasing on certain intervals, perhaps constant on other intervals. Show that the image integral

$$\phi[\mu](f) = \int_{-\infty}^{\infty} f(s)h(s)\, ds + \sum_{s^*} f(s^*)c(s^*) \tag{5.25}$$

is given by a density $h(s)$ and perhaps also some point masses $c(s^*)\delta_{s^*}$. Here

$$h(s) = \sum_{\phi(t)=s} w(t) \frac{1}{|\phi'(t)|} \tag{5.26}$$

and

$$c(s^*) = \int_{\phi(t)=s^*} w(t)\, dt. \tag{5.27}$$

5. What is the increasing right continuous function that defines the integral

$$\mu(g) = \int_{-\infty}^{\infty} g(x) \frac{1}{\pi} \frac{1}{1+x^2}\, dx \tag{5.28}$$

involving the Cauchy density?

6. What is the increasing function right continuous function that defines the δ_a integral given by $\delta_a(g) = g(a)$?

Chapter 6

Radon integrals

6.1 Radon measures

In the following we shall use the expression Radon integral to denote a particular kind of Lebesgue integral with special topological properties. This is more commonly called a Radon measure, even though it is defined on functions and not on sets. In view of the close connection between integrals and measures, it does not make much difference which terminology we use, and we shall most often use the expression Radon measure.

In the following X will be a locally compact Hausdorff space, but one can keep in mind the example $X = \mathbf{R}^n$. The space $C_c(X)$ is the space of all real continuous functions on X, such that each function in the space has compact support.

In this section we treat Radon measures. We shall only deal with positive Radon measures, though there is a generalization to signed Radon measures. So for our purposes, a Radon measure will be a positive linear function from $C_c(X)$ to $\mathbf{R}$.

In order to emphasize the duality between the space of measures and the space of continuous functions, we sometimes write the value of the Radon measure μ on the continuous function f as

$$\mu(f) = \langle \mu, f \rangle. \tag{6.1}$$

Theorem 6.1 *(Dini's theorem) Let $f_n \geq 0$ be a sequence of positive functions in $C_c(X)$ such that $f_n \downarrow 0$ pointwise. Then $f_n \to 0$ uniformly.*

Proof: The first function f_1 has compact support K. All the other functions also have support in K. Suppose that f_n does not converge to zero uniformly. Then there is an $\epsilon > 0$ such that for every N there exists an $n \geq N$ and an x such that $f_n(x) \geq \epsilon$. Since the f_n are monotone decreasing, this implies that there is an $\epsilon > 0$ such that for every N there exists an x with $f_N(x) \geq \epsilon$. Fix this $\epsilon > 0$. Let A_n be the set of all x such that $f_n(x) \geq \epsilon$. Each A_n is a closed

subset of K. Furthermore, the A_n are a decreasing family of sets. Finally, each A_n is non-empty. It follows from compactness that there is a point x in the intersection of the A_n. We conclude that for this x and for all n the inequality $f_n(x) \geq \epsilon$ is satisfied. So f_n does not converge to zero pointwise.

Theorem 6.2 *A Radon measure on $L = C_c(X)$ satisfies the monotone convergence theorem and therefore is an elementary integral.*

Proof: Consider a compact set K. Take a function g such that g is in $C_c(X)$ and $g = 1$ on K. Now consider an arbitrary function with support in K. Let M be the supremum of $|f|$. Then $-Mg \leq f \leq Mg$. It follows that $-M\mu(g) \leq \mu(f) \leq M\mu(g)$. In other words, $|\mu(f)| \leq \mu(g)M$. This shows that for functions f with support in K uniform convergence implies convergence of the integrals. The result then follows from Dini's theorem.

The importance of this theorem is that it gives us an easy way of constructing an integral. Take an arbitrary positive linear functional on the space of continuous functions with compact support. Then it uniquely defines an integral on the space of all measurable functions generated by the continuous functions with compact support. The reason this works is that the linear functional automatically satisfies the monotone convergence theorem, by the argument involving Dini's theorem. So it defines an elementary integral. By the general extension theorem of the last chapter, this elementary integral extends automatically to the smallest σ-algebra of functions that contains the continuous functions with compact support, and it continues to satisfy the monotone convergence theorem.

Example: Here is how to construct Radon measures defined for functions on the line. Let F be a right-continuous function that increases from m to M. Let G be the function that tries to be an inverse to F, as defined in the previous chapter. Thus

$$G(y) = \sup\{x \mid F(x) < y\}. \tag{6.2}$$

Define

$$\langle \mu, f \rangle = \int_m^M f(G(y))\, dy.$$

Then μ is a Radon measure. This Lebesgue-Stieltjes integral is often written in the form

$$\langle \mu, f \rangle = \int_{-\infty}^{\infty} f(x)\, dF(x).$$

The following examples will show why this is natural.

Example. Let $w(x) \geq 0$ be a locally integrable function on the line. Then

$$\langle \mu, f \rangle = \int_{-\infty}^{\infty} f(x)w(x)\, dx.$$

is a Radon measure. This is a special case of the first example. In fact, if F is the indefinite integral of w, then F is a continuous function. By making the change of variable $y = F(x)$ we recover the formula in the first example.

Example: Here is another special case of the first example. Let δ_a be the measure defined by

$$\langle \delta_a, f \rangle = f(a).$$

Then this is a Radon measure, the Dirac measure at the point a. This falls in the same framework. Let $F(x) = 0$ for $x < a$ and $F(x) = 1$ if $a \leq x$. Then $G(y) = a$ for $0 \leq a < 1$.

Example: Here is a somewhat more general situation that still falls in the framework of the first example. Let F be a function that is constant except for jumps of magnitude m_j at points a_j. Then $G(y) = a_j$ for $F(a_j) \leq y < F(a_j) + m_j$. Therefore,

$$\langle \mu, f \rangle = \sum_j m_j f(a_j),$$

so $\mu = \sum_j m_j \delta_{a_j}$.

6.2 Lower semicontinuous functions

Let us look more closely at the extension process in the case of a Radon measure. We begin with the positive linear functional on the space $L = C_c(X)$ of continuous functions with compact support. The construction of the integral associated with the Radon measure proceeds in the standard two stage process. The first stage is to consider the integral on the spaces $L\uparrow$ and $L\downarrow$. The second stage is to use this extended integral to define the integral of an arbitrary summable function.

A function f from X to $(-\infty, +\infty]$ is said to be lower semicontinuous (LSC) if for each real a the set $\{x \mid f(x) > a\}$ is an open set. A function f from X to $[-\infty, +\infty)$ is said to be upper semicontinuous (USC) if for each real a the set $\{x \mid f(x) < a\}$ is an open set. Clearly a continuous real function is both LSC and USC.

Theorem 6.3 *If each f_n is LSC and if $f_n \uparrow f$, then f is LSC. If each f_n is USC and if $f_n \downarrow f$, then f is USC.*

It follows from this theorem that space $L\uparrow$ consists of functions that are LSC. Similarly, the space $L\downarrow$ consists of functions that are USC. These functions can already be very complicated. The first stage of the construction of the integral is to use the monotone convergence theorem to define the integral on the spaces $L\uparrow$ and $L\downarrow$.

In order to define the integral for a measurable functions, we approximate such a function from above by a function in $L\uparrow$ and from below by a function in $L\downarrow$. This is the second stage of the construction. The details were presented in the last chapter.

The following is a useful result that we state without proof.

Theorem 6.4 *If μ is a Radon measure and if $1 \leq p < \infty$, then $C_c(X)$ is dense in $L^p(X, \mu)$.*

Notice carefully that the corresponding result for $p = \infty$ is false. The uniform closure of $C_c(X)$ is $C_0(X)$, which in general is much smaller than $L^\infty(X, \mu)$. A bounded function does not have to be continuous, nor does it have to vanish at infinity.

6.3 Weak* convergence

As before, we consider only positive Radon measures, though there is a generalization to signed Radon measures. We consider finite Radon measures, that is, Radon measures for which $\langle \mu, 1\rangle < \infty$. Such a measure extends by continuity to $C_0(X)$, the space of real continuous functions that vanish at infinity. In the case when $\langle \mu, 1\rangle = 1$ we are in the realm of probability.

In this section we describe *weak* convergence* for Radon measures. In probability this is often called *vague* convergence. A sequence μ_n of finite Radon measures is said to weak* converge to a Radon measure μ if for each f in $C_0(X)$ the numbers $\langle \mu_n, f\rangle \to \langle \mu, f\rangle$.

The importance of weak* convergence is that it gives a sense in which two probability measures with very different qualitative properties can be close. For instance, consider the measure

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{\frac{k}{n}}.$$

This is a Riemann sum measure. Also, consider the measure

$$\langle \lambda, f\rangle = \int_0^1 f(x)\, dx.$$

This is Lebesgue measure on the unit interval. Then $\mu_n \to \lambda$ in the weak* sense, even though each μ_n is discrete and λ is continuous.

A weak* convergent sequence can lose mass. For instance, a sequence of probability measures μ_n can converge in the weak* sense to zero. A simple example is the sequence δ_n. The following theory shows that a weak* convergent sequence cannot gain mass.

Theorem 6.5 *If $\mu_n \to \mu$ in the weak* sense, then $\langle \mu, f\rangle \le \lim_n \inf_{k \ge n} \langle \mu_k, f\rangle$ for all $f \ge 0$ in $BC(X)$.*

Proof: It is sufficient to show this for f in BC with $0 \le f \le 1$. Choose $\epsilon > 0$. Let $0 \le g \le 1$ be in C_0 so that $\langle \mu, (1-g)\rangle < \epsilon$. Notice that gf is in C_0. Furthermore, $(1-g)f \le (1-g)$ and $gf \le f$. It follows that

$$\langle \mu, f\rangle \le \langle \mu, gf\rangle + \langle \mu, (1-g)\rangle \le \langle \mu, gf\rangle + \epsilon \le \langle \mu_k, gf\rangle + 2\epsilon \le \langle \mu_k, f\rangle + 2\epsilon \quad (6.3)$$

for k sufficiently large.

The following theorem shows that if a weak* convergent sequence does not lose mass, then the convergence extends to all bounded continuous functions.

Theorem 6.6 *If $\mu_n \to \mu$ in the weak* sense, and if $\langle \mu_n, 1\rangle \to \langle \mu, 1\rangle$, then $\langle \mu_n, f\rangle \to \langle \mu, f\rangle$ for all f in $BC(X)$.*

Proof: It is sufficient to prove the result for f in BC with $0 \leq f \leq 1$. The preceding result gives an inequality in one direction, so it is sufficient to prove the inequality in the other direction. Choose $\epsilon > 0$. Let $0 \leq g \leq 1$ be in C_0 so that $\langle \mu, (1-g)\rangle < \epsilon$. Notice that gf is in C_0. Furthermore, $(1-g)f \leq (1-g)$ and $gf \leq f$. For this direction we note that the extra assumption implies that $\langle \mu_n, (1-g)\rangle \to \langle \mu, (1-g)\rangle$. We obtain

$$\langle \mu_n, f\rangle \leq \langle \mu_n, gf\rangle + \langle \mu_n, (1-g)\rangle \leq \langle \mu, gf\rangle + \langle \mu, (1-g)\rangle + 2\epsilon \leq \langle \mu, gf\rangle + 3\epsilon \leq \langle \mu, f\rangle + 3\epsilon \tag{6.4}$$

for n sufficiently large.

It is not true in general that the convergence works for discontinuous functions. Take the function $f(x) = 1$ for $x \leq 0$ and $f(x) = 0$ for $x > 0$. Then the measures $\delta_{\frac{1}{n}} \to \delta_0$ in the weak* sense. However $\langle \delta_{\frac{1}{n}}, f\rangle = 0$ for each n, while $\langle \delta_0, f\rangle = 1$.

We now want to argue that the convergence takes place also for certain discontinuous functions. A quick way to such a result is through the following concept. For present purposes, we say that a bounded measurable function g has μ-negligible discontinuities if for every $\epsilon > 0$ there are bounded continuous functions f and h with $f \leq g \leq h$ and such that $\mu(f)$ and $\mu(h)$ differ by less than ϵ.

Example: If λ is Lebesgue measure on the line, then every piecewise continuous function with jump discontinuities has λ-negligible discontinuities.

Example: If δ_0 is the Dirac mass at zero, then the indicator function of the interval $(-\infty, 0]$ does not have δ_0-negligible discontinuities.

Theorem 6.7 *If $\mu_n \to \mu$ in the weak* sense, and if $\langle \mu_n, 1\rangle \to \langle \mu, 1\rangle$, then $\langle \mu_n, g\rangle \to \langle \mu, f\rangle$ for all bounded measurable g with μ-negligible discontinuities.*

Proof: Take $\epsilon > 0$. Take f and h in BC such that $f \leq g \leq h$ and $\mu(f)$ and $\mu(h)$ differ by at most ϵ. Then $\mu_n(f) \leq \mu_n(g) \leq \mu_n(h)$ for each n. It follows that $\mu(f) \leq \lim_n \inf_{k \geq n} \mu_n(g) \leq \lim_n \sup_{k \geq n} \mu_n(g) \leq \mu(h)$. But also $\mu(f) \leq \mu(g) \leq \mu(h)$. This says that $\lim_n \inf_{k \geq n} \mu_n(g)$ and $\lim_n \sup_{k \geq n} \mu_n(g)$ are each within ϵ of $\mu(g)$. Since $\epsilon > 0$ is arbitrary, this proves that $\lim_n \mu_n(g)$ is $\mu(g)$.

6.4 The central limit theorem

Let X be a random variable with mean μ. Then the *centered* X is the random variable $X - \mu$. It measures the deviation of X from its expected value. Let X be non-constant with standard deviation $\sigma > 0$. The *standardized* X is the random variable

$$Z = \frac{X - \mu}{\sigma}. \tag{6.5}$$

It measures the deviation of X from its expected value in units of the standard deviation.

A *normal* (or *Gaussian*) random variable X with mean μ and variance σ^2 is a random variable whose distribution has density

$$\rho_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{6.6}$$

The centered version of X will have density

$$\rho_{X-\mu}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \tag{6.7}$$

The standardized version Z of X will have density

$$\rho_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{6.8}$$

We shall most often work with the centered or standardized versions. The tables of the distributions of normal random variables are for the standardized version.

Why is the normal distribution so important? The explanation traces back to the following remarkable property of the normal distribution.

Theorem 6.8 *Let X and Y be independent centered normal random variables with variances σ^2. Let $a^2 + b^2 = 1$. Then $aX + bY$ is a centered normal random variable with variance σ^2.*

Proof: The joint density of X and Y is

$$\rho(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}. \tag{6.9}$$

We can use properties of exponentials to write this as

$$\rho(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{6.10}$$

Let $Z = aX + bY$ and $W = -bX + aY$. We can write the joint density of X and Y as

$$\rho(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(ax+by)^2+(-bx+ay)^2}{2\sigma^2}}. \tag{6.11}$$

Now we can factor this again as

$$\rho(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(ax+by)^2}{2\sigma^2}} e^{-\frac{(-bx+ay)^2}{2\sigma^2}}. \tag{6.12}$$

Let $Z = aX + bY$ and $W = -bX + aY$. The transformation $z = ax + by$, $w = -bx + ay$ is a rotation. Therefore it preserves areas: $dz\,dw = dx\,dy$. It follows that Z and W have joint density

$$\rho(z, w) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{6.13}$$

Corollary 6.1 *Let $X_1, \ldots, X_n$ be independent random variables with centered normal distributions with variance σ^2. Then the sum $X_1 + X_2 + \cdots + X_n$ is also normal, with variance $n\sigma^2$.*

Proof: We can prove this by showing that it is true when $n = 1$ and by showing that for all n, if it is true for n, it is also true for $n + 1$.

The fact that it is true for $n = 1$ is obvious. Suppose that it is true for n. Then the sum $X_1 + \cdots + X_n$ is normal with variance $n\sigma^2$. It follows that $(X_1 + \cdots + X_n)/\sqrt{n}$ is normal with variance σ^2. The next term X_{n+1} is also normal with variance σ^2. Take $a = \sqrt{n}/\sqrt{n+1}$ and $b = 1/\sqrt{n+1}$. Then by the theorem $(X_1 + \cdots + X_n + X_{n+1})/\sqrt{n+1}$ is normal with variance σ^2. It follows that $X_1 + \cdots + X_{n+1}$ is normal with variance $(n+1)\sigma^2$.

The next theorem is the famous central limit theorem. It is a spectacular result. It says that the deviations of sums of independent random variables with finite standard deviation have a distribution that is universal. No matter what distribution we start with, when we have such a sum with a large number of terms, the distribution is normal.

Theorem 6.9 *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and standard deviation $\sigma < \infty$. Let $\bar{X}_n$ be their sample mean. Then the distribution of the standardized variable*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{6.14}$$

converges in the weak sense to the standard normal distribution.*

Recall that weak* convergence for a sequence of probability measures converging to a probability measure says that for each bounded continuous function f the expectation $E[f(Z_n)]$ converges to $E[f(Z)]$ as $n \to \infty$. In the central limit theorem the Z is a standard normal random variable, so

$$E[f(Z)] = \int_{-\infty}^{\infty} f(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz. \tag{6.15}$$

It is well to keep in mind that since the limiting normal distribution has a density, the convergence also extends to bounded continuous functions with jump discontinuities. In particular, the probabilities associated with intervals converge.

Proof: By centering, the proof is reduced to the following result. Let $X_1, X_2, X_3, \ldots$ be independent identically distributed random variables with mean zero and variance σ^2. Let Z be a normal random variable with mean zero and variance σ^2. Let f be a bounded continuous function. Then

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] \to E[f(Z)] \tag{6.16}$$

as $n \to \infty$.

We shall prove the theorem for the case when f is a smooth bounded function with bounded derivatives. The case of bounded continuous functions follows by an approximation argument.

The idea is to compare the sequence $X_1, \ldots, X_n$ with a corresponding sequence $Z_1, \ldots, Z_n$ of normal random variables. The strategy is to do the comparison in n stages. First compare $Z_1, Z_2, Z_2 \ldots, Z_n$ to $X_1, Z_2, Z_3 \ldots, Z_n$. Then compare $X_1, Z_2, Z_3 \ldots, Z_n$ to $X_1, X_2, Z_3, \ldots, Z_n$, and so on. Each time only one variable is changed from normal to another form. The effect of this change is very small, so the total effect is small. On the other hand, for the sequence $Z_1, \ldots, Z_n$ we already have the central limit theorem, by explicit computation.

To make this explicit, write

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] - E[f(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}})] = \sum_{i=1}^{n} (E[f(W_i + \frac{X_i}{\sqrt{n}})] - E[f(W_i + \frac{Z_i}{\sqrt{n}})]). \tag{6.17}$$

Here

$$W_i = \frac{X_1 + \cdots + X_{i-1} + Z_{i+1} + \cdots + Z_n}{\sqrt{n}}. \tag{6.18}$$

Next write

$$f(W_i + \frac{X_i}{\sqrt{n}}) = f(W_i) + f'(W_i)\frac{X_i}{\sqrt{n}} + \frac{1}{2} f''(W_i)\frac{X_i^2}{n} + R_i. \tag{6.19}$$

Here R_1 is a remainder. In the same way, write

$$f(W_i + \frac{Z_i}{\sqrt{n}}) = f(W_i) + f'(W_i)\frac{Z_i}{\sqrt{n}} + \frac{1}{2} f''(W_i)\frac{Z_i^2}{n} + S_i. \tag{6.20}$$

Again S_i is a remainder.

Now calculate the expectations. The expectations of the zeroth order terms are the same.

Use independence to calculate

$$E[f'(W_i)\frac{X_i}{\sqrt{n}}] = E[f'(W_i)]E[\frac{X_i}{\sqrt{n}}] = 0. \tag{6.21}$$

and similarly for the other case. The expectations of the first order terms are zero.

Use independence to calculate

$$E[f'(W_i)\frac{X_i^2}{n}] = E[f'(W_i)]E[\frac{X_i^2}{n}] = E[f''(W_i)]\frac{\sigma^2}{n} \tag{6.22}$$

and similarly for the other case. The expectations of the second order terms are the same.

This is the heart of the argument: up to second order everything cancels exactly. This shows that the sum of interest is a sum of remainder terms.

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] - E[f(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}})] = \sum_{i=1}^{n} (E[R_i] - E[S_i]). \tag{6.23}$$

The only remaining task is to show that these remainder terms are small. This is a somewhat technical task. Here is an outline of how it goes.

Write

$$R_i = \frac{1}{2}(f''(W_i + \alpha \frac{X_i}{\sqrt{n}}) - f''(W_i))\frac{X_i^2}{n}. \tag{6.24}$$

In the same way, write

$$S_i = \frac{1}{2}(f''(W_i + \beta \frac{Z_i}{\sqrt{n}}) - f''(W_i))\frac{Z_i^2}{n}. \tag{6.25}$$

Look at the first remainder $E[R_i]$. Break up the sample space into two parts. The first part is where $|X_i| \leq \epsilon\sqrt{n}$. The second part is where $|X_i| > \epsilon\sqrt{n}$.

Assume that the absolute value of the third derivative of f is bounded by a constant C. Then the size of the first part of the first remainder is bounded by

$$E[|R_i|; |X_i| \leq \epsilon\sqrt{n}] \leq \frac{1}{2}C\epsilon\frac{\sigma^2}{n}. \tag{6.26}$$

Assume that the absolute value of the second derivative of f is bounded by a constant M. The size of the second part of the first remainder is bounded by

$$E[|R_i|; |X_i| > \epsilon\sqrt{n}] \leq M\frac{1}{n}E[X_i^2; X_i > \epsilon\sqrt{n}]. \tag{6.27}$$

Let ϵ depend on n, so we can write it as a sequence ϵ_n. Then the first remainder is bounded by

$$E[|R_i|] \leq \frac{1}{2}C\epsilon_n\frac{\sigma^2}{n} + M\frac{1}{n}E[X_i^2; X_i > \epsilon_n\sqrt{n}]. \tag{6.28}$$

Thus the absolute value of the first sum is bounded by

$$\sum_{i=1}^{n} E[|R_i|] \leq \frac{1}{2}C\epsilon_n\sigma^2 + ME[X_i^2; X_i > \epsilon_n\sqrt{n}]. \tag{6.29}$$

Take $\epsilon_n \to 0$ with $\epsilon_n\sqrt{n} \to \infty$. Then the first sum goes to zero as $n \to \infty$. The same idea works for the other sum. This completes the proof.

Corollary 6.2 *Let $E_1, \ldots, E_n$ be independent events, each with probability p. Let N_n be the number of events that happen, and let $f_n = N_n/n$ be the relatively frequency of the events that happen. Then distribution of the standardized variable*

$$Z_n = \frac{N_n - np}{\sqrt{n}\sqrt{p(1-p)}} = \frac{f_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \tag{6.30}$$

converges in the weak sense to the standard normal distribution.*

The preceding corollary is the first form of the central limit theorem that was discovered. It illustrates the utility of the notion of weak* convergence. Each Z_n has a discrete distribution, but the limiting Z is normal and hence has a continuous distribution.

6.5 Problems

1. Show that $f : X \to (-\infty, +\infty]$ is lower semicontinuous (LSC) if and only if $x_n \to x$ implies $\lim_n \inf_{k \geq n} f(x_k) \geq f(x)$.

2. Show that if $f_n \uparrow f$ pointwise and each f_n is LSC, then f is LSC.

3. Show that if $f : X \to (-\infty, +\infty]$ is LSC, and X is compact, then there is a point in X at which f has a minimum value.

4. Show by example that if $f : X \to (-\infty, +\infty]$ is LSC, and X is compact, then there need not be a point in X at which X has a maximum value.

5. Give an example of a function $f : [0, 1] \to (-\infty, +\infty]$ that is LSC but is discontinuous at every point.