Probability Theory: A Summary

ZWEISTEIN

Desu-Cartes

Contents

1 Measure Theory

1.1 Introduction

In order to define modern probability, we will have to start by making a necessary detour in measure theory and integration. In general, we will not define a probability measure on all subsets of our space. Instead, perhaps in analogy to topology and open sets, we define which sets we decide to work with. Those sets will form a particular algebraic structure, and we will define a probability on it. After that, we define integration in general and consider the important question of interchanging limits between our integral and other analytical objects such as sequences, series and other integrals. With this out of the way, we can define function spaces and look at what kind of structure our functions form. Finally, we can define probability theory using all the tools we've built so far, and analyze what kind of modes of convergence exist between probabilistic objects. This section culminates in the strong law of large numbers which reinforces our intuition about what probability really is about.

1.2 Measures and algebras

DEFINITION 1.1. We consider a universe of events (a set really) Ω. Let $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ be a family of sets over Ω. We say that $\mathcal{F}$ is a *σ-algebra* if the following properties are verified:

1. For any countable family of subsets in $\mathcal{F}$, their union is in $\mathcal{F}$.
2. For any subset in $\mathcal{F}$, their complement is in $\mathcal{F}$.
3. The universe Ω is in $\mathcal{F}$.

It is clear from this definition that countable intersections of elements of $\mathcal{F}$ are in $\mathcal{F}$ and that the empty set is in $\mathcal{F}$. We call elements $A \in \mathcal{F}$ *events* or more generally, *measurable sets*. The pair $(\Omega, \mathcal{F})$ is called a measurable space.

In some cases, we can take $\mathcal{P}(\Omega)$ as our σ-algebra. We will usually consider this σ-algebra when working over countable spaces such as $\mathbb{N}$. Since $\mathcal{P}(\Omega)$ is always a σ-algebra for any universe Ω, we can use this fact to understand the notion of σ-algebra generated by a subset of Ω.

DEFINITION 1.2. Let $X \in \mathcal{P}(\Omega)$. Then the *$\sigma$-algebra generated by X* is the intersection of all σ-algebras over Ω that contain X. It is clear this intersection exists from the discussion above. We will denote this σ-algebra by $\sigma(X)$.

It is not so clear why we need σ-algebras initially. Surely, if $\mathcal{P}(\Omega)$ always works, why not just take that at all times? Choosing which sets are measurable will fine-tune the notion of "measurability" itself, just like specifying a topology on X will restrict topological notions such as convergence or continuity. Just as in topology, we also have special maps that preserve the notion of being measurable.

DEFINITION 1.3. Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. We say a function $f : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ is *measurable* if it respects the σ-algebra structure. In other words, for any $B \in \mathcal{F}'$, we have that $f^{-1}(B) \in \mathcal{F}$. If $\mathcal{F} = \sigma(X)$, then it is sufficient to check this for any subset of X instead. From now on, we write $f : \Omega \to \Omega'$ instead of the more cumbersome notation above.

This makes clearer the notion of "measurability" and how similar it is to topology. In topology, open sets specify which functions are continuous. In measurable spaces, elements of σ-algebra specify which functions are measurable.

Proposition 1.4. *Measurable functions respect the usual algebraic operations. Sums, (scalar and pointwise) products, quotients and compositions of measurable functions are measurable where it makes sense. Furthermore, if the target space of the function is a metric space, pointwise limits of measurable functions are measurable. Supremums, infimums and their limits of real-valued measurable functions are measurable as well.*

Proof. The proofs are either not very illuminating or extremely easy. We just show composition. Suppose $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to \Omega_3$ are measurable functions whose domains and codomains match as one would want. Then the preimage of $g \circ f$ is $f^{-1}(g^{-1}(A))$. Suppose $A \in \mathcal{F}_3$. Then $B = g^{-1}(A) \in \mathcal{F}_2$. In turn, $C = f^{-1}(B) \in \mathcal{F}_1$. This holds because both f and g are measurable. Thus for $A \in \mathcal{F}_3$, $(g \circ f)(A) \in \mathcal{F}_1$ which shows $g \circ f$ is measurable. □

As one can see, the class of measurable functions is quite rich. This is a good thing. Since we will define integration on those functions, we want as many functions as possible to be at least measurable. It is obvious that continuous functions are measurable provided that the σ-algebra on the domain is the one generated by open (or closed) sets. We will call this σ-algebra the *Borel* algebra. In the case of the real line, we will write it as $\mathcal{B}(\mathbb{R})$. We now define the main object of study in measure theory.

DEFINITION 1.5. A *measure* is a function $\mu : \mathcal{F} \to \mathbb{R}^+ \cup \{\infty\}$ such that the following holds:

1. $\mu(\varnothing) = 0$
2. $\mu(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} \mu(X_i)$, for any countable family of disjoint sets $(X_i)_{i \in I}$.

The second property is called σ-additivity (sometimes also known as countable additivity). If $\mu(\Omega) = 1$, we say that μ is a *probability measure*, or more simply, a *probability*. The triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. If $\mu = \mathbb{P}$ is a probability, then we call the space $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space*, and a measurable function on it a *random variable*.

Proposition 1.6 (Properties of measures)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $A, B \in \mathcal{F}$. Then:*

1. $\mu(A) = \mu(A \setminus B) + \mu(A \cap B)$;

2. *If $\mu(A) < \infty$, then $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$;*

3. *If $A \subseteq B$, then $\mu(A) \leq \mu(B)$. If in addition, $\mu(A) < \infty$, then $\mu(B \setminus A) = \mu(B) - \mu(A)$;*

4. *If $(A_n)_n$ is an increasing sequence and for every n, $A_n \in \mathcal{F}$, then $\mu(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$. If the sequence is decreasing and $\mu(A_1) < \infty$, then $\mu(\bigcap_{i=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$. Furthermore, the limit converge increasingly (resp. decreasingly).*

Proof. Those are all easy properties that we leave to the reader. For 4, use a decomposition of your space that allows you to apply σ-additivity. □

It will be important for us to define measures based on random variables. Indeed, suppose $\mathbb{P}$ is a probability. We will often be interested in understanding the probability related to values of random variables. Suppose X is a real-valued random variable on Ω (the space, by abuse of notation). We are interested in understanding $\mathbb{P}(\omega : X(\omega) \in B)$ for an event B. We denote this function $\mathbb{P}_X(B)$. It is not too hard to see that this is a measure on $\Omega' = \mathbb{R}$. This will become important in the probability section.

Notice that we haven't actually talked about the problem of existence of measures. Indeed, we will not concern ourselves with it. For now, we just assume there's a magic theorem (a few, actually) that allows us to construct all the measures we need. We finish this section by presenting two important measures.

Proposition 1.7 (Lebesgue measure)**.** *There exists a translation-invariant measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda([a, b]) = b - a$.*

Proposition 1.8 (Product measure)**.** *Suppose Ω_1, Ω_2 are two measure spaces with measures m_1 (resp. m_2). Then there exists a measure $m_1 \otimes m_2$ on $\Omega_1 \times \Omega_2$ such that for all $A \in \Omega_1$, $B \in \Omega_2$, we have $(m_1 \otimes m_2)(A \times B) = m_1(A)m_2(B)$.*

2 Integration Theory

2.1 Introduction

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We wish to define a linear functional $f \longmapsto \int_\Omega f d\mu$ on the space of real-valued measurable functions. When we learn about the Darboux-Riemann integral, we start by subdividing the domain of integration. Here, we take a different approach. In a sense, we subdivide the range of the function and then sum. This subdivision comes in the form of limits of finite sums of indicator functions, which we define next.

Definition 2.1. The *indicator function* of a set A is the function 1_A which is equal to 1 on A, 0 everywhere else. It is clear that 1_A is measurable if and only if A is. A simple function is a finite sum of indicator functions.

We proceed in three steps to define our integral. First we construct our integral for indicator functions only. Then we define it for simple functions. Finally, we define it for positive functions and extend that definition to measurable functions in general.

2.2 Integration of real-valued functions

We now construct the integral for each class of functions discussed above. In this subsection, a measurable function means a real-valued measurable function.

DEFINITION 2.2 (Existence of the integral). Suppose f is a measurable function on $(\Omega, \mathcal{F}, \mu)$.

1. If $f = 1_A$ is an indicator function, then we define $\int_\Omega 1_A \, d\mu = \int_A d\mu = \mu(A)$. This exists because f is supposed measurable.
2. If $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ is a simple function, then we define $\int_\Omega f d\mu = \sum_{i=1}^n \alpha_i \int_\Omega A_i \, d\mu$. This exists because of 1.
3. If f is positive, we define $\int_\Omega f d\mu = \sup\left\{\int_\Omega g d\mu\right\}$, where g is a simple function and $g \leq f$. More general, for a measurable function we decompose it as $f = f^+ - f^-$ and apply 3 to both terms. Notice that our integral might be undefined here.

As we have seen in the third part of the proposition, it is quite possible that for a general measurable function, its integral is undefined. To fix that, we introduce a new class of functions, called *integrable functions*.

DEFINITION 2.3. Let f be a measurable function. If $\int_\Omega |f| d\mu < \infty$, we say that the function is integrable. The class of integrable functions is a vector space denoted by $\mathcal{L}^1(\mu)$. We will look more into this space and other similar spaces in section 3.

DEFINITION 2.4 (Almost-everywhere convergence). We say that two functions are *equal almost everyhwere* if $f = g$ except maybe on a set of measure zero. A sequence of functions (f_n) converges to f almost everywhere (a.e., or sometimes almost surely, a.s.) if it converges pointwise to f on the complement of a set of measure zero. Finally, we say that $P(\omega)$ is a property almost-everywhere if P holds everywhere except possibly on a set of measure zero. For instance, for almost all x means for all x except possibly for those in a set N of measure zero.

It is important to notice that almost-everywhere properties will pop up a lot in Lebesgue integration. This is so because our integral doesn't detect this property. If $f = g$ almost everywhere in $\mathcal{L}^1(\mu)$, then $\int_\Omega f - g \, d\mu = 0$ always, as seen in the next proposition.

Proposition 2.5 (Properties of the integral)**.** *Suppose f is a measurable function.*

1. *The operator $\mathcal{I}(f)$ that sends a measurable function to its integral is a linear operator from the space of real-valued measurable functions to $[0, +\infty]$;*

2. *If $f \leq g$ are both in $\mathcal{L}^1(\mu)$, then $\int_\Omega f\,d\mu \leq \int_\Omega g\,d\mu$. In general, if the functions are positive, the same holds true.*

3. *If $f \in \mathcal{L}^1(\mu)$, then $\left|\int_\Omega f\,d\mu\right| \leq \int_\Omega |f|\,d\mu$.*

4. *If A, B are disjoint measurable sets, then $\int_{A\cup B} f\,d\mu = \int_A f\,d\mu + \int_B f\,d\mu$. If A has measure zero, then $\int_A f\,d\mu = 0$ always. If $f = g$ a.e.,then $\int_\Omega f\,d\mu = \int_\Omega g\,d\mu$.*

Proof. For most of these properties, start by considering them first for indicator functions, then for simple functions, and then apply Theorem 2.7. For 4, decompose $1_{A\cup B}$ and notice that the measure of N is zero, which will make the integral zero. □

Right now, it is not so clear why this integral is more useful or powerful than the usual Darboux integral. While there are other forms of integration on $\mathbb{R}$ such as the gauge integral, the Lebesgue integral is extremely well-suited to analysis for two reasons. The first one is because of convergence theorems, the second one is the main topic of section 3. We now turn to those convergence theorems. This is the crux of Lebesgue integration theory.

Theorem 2.6 (Fatou's Lemma)**.** *For any sequence of positive measurable functions (f_n), we have*

$$\int_\Omega \liminf_{n\to\infty} f_n\,d\mu \leq \liminf_{n\to\infty} \int_\Omega f_n\,d\mu.$$

Proof. This inequality comes from the fact that for $k \geq 1$,

$$\int \left(\inf_{n\geq k} f_n\right) d\mu \leq \inf_{n\geq k} \int f_n\,d\mu$$

because the integral is increasing. Letting $k \to \infty$ and using the definition of $\liminf_{n\to\infty}$ and properties of the integral above does the job. □

Don't be fooled by the name, this is a powerful result in its own right. This actually is a lemma though, because it is used in the proof of the dominated convergence theorem we will soon see.

Theorem 2.7 (Monotone convergence theorem)**.** *For any positive increasing sequence of functions (f_n) with $\lim_{n\to\infty} f_n = f$, we have that*

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

Proof. Since (f_n) is increasing, the integral is as well and thus the limit $r = \lim_{n\to\infty} \int_\Omega f_n \, d\mu$ exists in $[0, +\infty]$. This shows $r \leq \int_\Omega f \, d\mu$. We just need to show the other inequality now. To do this, fix c in $[0, 1)$ and g a simple function such that $g \leq f$. Then, according to the monotonicity of the integral,

$$\int_\Omega f_n \, d\mu \geq \int_\Omega 1_{\{f_n \geq cg\}} f_n \, d\mu \geq \int_\Omega 1_{\{f_n \geq cg\}} g \, d\mu.$$

Writing out g gives us the inequality

$$\int_\Omega 1_{\{f_n \geq cg\}} g \, d\mu = \sum_{x \in g(\Omega)} \mu\left(\{g = x\} \cap \{f_n \geq cx\}\right) x.$$

Using the monotone property of measures, we get as $n \to \infty$

$$\sum_{x \in g(\Omega)} \mu\left(\{g = x\} \cap \{f_n \geq cx\}\right) x. = \sum_{x \in g(\Omega)} \mu\left(\{g = x\}\right) x = \int_\Omega g \, d\mu.$$

Thus $r \geq c \int_\Omega g \, d\mu$, and since both $c \in [0, 1)$ and g are arbitrary, we are done. □

A very handy theorem, this is not true at all in the case of Riemann integration. Finally, we get to maybe the most powerful theorem in the theory, the celebrated dominated convergence theorem.

Theorem 2.8 (Dominated Convergence Theorem)**.** *Let (f_n) be a sequence of real-valued measurable functions. If we have that:*

1. *f_n converges almost everywhere to a function f;*
2. *There exists $h \in \mathcal{L}^1(\mu)$ such that $|f_n(\omega)| \leq h(\omega)$, for all n;*

then

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

Proof. Let $(f_n)_{n\geq 1}$ be a convergent sequence of measurable functions dominated by an integrable function h. The measurable functions $f_n - g$ (resp. $f_n + g$) are thus positive and tend to $g - f$ (resp. $g + f$) as $n \to \infty$. We can thus apply Fatou's lemma to get

$$\int_\Omega (g \pm f) \, d\mu \leq \liminf_{n\to\infty} \int_\Omega (g \pm f) \, d\mu.$$

Since g is integrable, we can subtract it from both sides to get

$$\liminf_{n\to\infty} \int_\Omega f_n \, d\mu \geq \int_\Omega f \, d\mu.$$

But notice now that

$$\limsup_{n\to\infty} \int_\Omega f_n \, d\mu \leq \int_\Omega f \, d\mu$$

which gives us the desired result. □

There exists a generalized version of this theorem where each f_n is bounded by a specific g_n with extra conditions but we will not need it here.

2.3 Multiple Integrals and Derivatives

In this subsection, we propose to make the theorems we've developed in the preceding section to good use. Let U be an open subset of a metric space E, and Ω be a measure space. We can use the dominated convergence theorem to analyze the relationship between integrals, continuity and differentiability. More specifically, we are interested in functions of the form

$$F(x) = \int_\Omega f(x,\omega)\, d\mu(\omega)$$

where $f(x,\omega)$ is a function from $U \times \Omega \to \mathbb{R}$.

Theorem 2.9 (Continuity under the integral sign)**.** *Suppose that:*

1. *For all $x \in U$, the function $w \mapsto f(x,w)$ is measurable;*
2. *For almost all $\omega \in \Omega$, the function $x \mapsto f(x,\omega)$ is continuous at $x_0 \in U$;*
3. *There exists $h \in \mathcal{L}^1(\mu)$ such that $|f(x,\omega)| \leq h(\omega)$ almost surely.*

Then $F(x)$ as defined above is well-defined and continuous at x_0.

Proof. We get the existence from property 3. To show continuity, it's enough to show that $F(x_n) \to F(x_0)$ for any sequence in U such that $x_n \to x_0$. Let (x_n) be such a sequence. Then, from property 2 we have that $f(x_n,\omega) \to f(x_0,\omega)$. From 3 and the DCT, we get $F(x_n) \to F(x_0)$. □

Theorem 2.10 (Differentiation under the integral sign)**.** *Here we assume that U is an open interval in $\mathbb{R}$. Suppose that:*

1. *For all $x \in U$, the function $w \mapsto f(x,w)$ is integrable;*
2. *For almost all $\omega \in \Omega$, and for all $x \in U$, the partial derivative $\frac{\partial f}{\partial x}(x,\omega)$ exists and verifies*
$$\left|\frac{\partial f}{\partial x}(x,\omega)\right| \leq h(\omega)$$
where $h \in \mathcal{L}^1(\mu)$.

Then the function defined above is well-defined and differentiable. Furthermore, we have that

$$F'(x) = \int_\Omega \frac{\partial f}{\partial x}(x,\omega)\, d\mu(\omega).$$

Proof. Let (x_n) be a sequence in U converging to $x \in U$. Then $g_n(\omega) = \frac{f(x_n,\omega)-f(x,\omega)}{x_n - x}$ converges almost everywhere to $\frac{\partial f}{\partial x}(x,\omega)$. Applying the mean value theorem, we get

$$|g_n(\omega)| \leq \sup_{0\leq\theta\leq 1} \left|\frac{\partial f}{\partial x}\left(\theta x + (1-\theta)x_n, \omega\right)\right| \leq h(\omega).$$

We now use the DCT to conclude that

$$\lim_{n\to\infty} \frac{F_n(x) - F(x)}{x_n - x} = \lim_{n\to\infty} \int_\Omega g_n(\omega)\, d\mu(\omega) = \int_\Omega \frac{\partial f}{\partial x}(x,\omega)\, d\mu(\omega).$$

□

These two theorems help us determine when we can do the naive differentiation under the integral sign. We see here just why the dominated convergence theorem is useful. Domination is a hypothesis in both theorems. We now turn to the issue of integration over product spaces. Recall that if $(\Omega_1, \mathcal{A}_1, m_1)$ and $(\Omega_2, \mathcal{A}_2, m_2)$ are two measure spaces, we can look at the measure space $(\Omega_1 \times \Omega_2, m_1 \otimes m_2)$. One particular question we want to answer is when is it true that integration over the product space is akin to integrating twice over the respective measure spaces. For this, we will need a certain hypothesis.

Definition 2.11. We say that a measure is σ-finite if there exists a countable family of measurable sets $(A_i)_{i\in I}$ such that $\Omega = \bigcup_{i\in I}^{\infty}$ and $\mu(A_i) < \infty$, for all $i \in \mathbb{N}$.

Proposition 2.12 (Measurability of slices)**.** *Suppose $f(x_1, x_2)$ is measurable with respect to the product σ-algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$. Then the slice function $x_2 \mapsto f(x_1, x_2)$ is measurable with respect to $\mathcal{A}_2$. Likewise, $x_1 \mapsto f(x_1, x_2)$ is measurable with respect to $\mathcal{A}_1$.*

Proof. Apply the usual machinery starting from indicator functions, then simple functions and use the MCT to prove the property for positive functions. □

We are now in a position to state the theorem that will allow us to interchange multiple integrals. This is the famous Fubini-Tonelli theorem.

Theorem 2.13 (Fubini)**.** *Let m_1, m_2 be two σ-finite measures. Then:*

1. *If $f(x_1, x_2)$ is positive and measurable with respect to the product measure $m_1 \otimes m_2$, then*

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f(x_1,x_2)\, d(m_1 \otimes m_2) &= \int_{\Omega_1} \left(\int_{\Omega_2} f(x_1,x_2)\, dm_2(x_2) \right) dm_1(x_1) \\ &= \int_{\Omega_2} \left(\int_{\Omega_1} f(x_1,x_2)\, dm_1(x_1) \right) dm_2(x_2)\end{aligned}$$

2. *If $f(x_1, x_2)$ isn't positive but is integrable, the same property holds true.*

Proof. We mentioned earlier a few magic theorems that let us assume certain measures exist. In our case this was both Carathéodory's extension theorem, and the $\lambda - \pi$ theorem. While we not delve into those here, we just mention that the second one is needed for this proof. The idea is that you consider the three members of the equality above and apply the usual machinery. First you consider indicator functions, then you use the $\lambda - \pi$ theorem to extend that to simple functions, and you apply the MCT as usual. □

2.4 Change of Variables

Right after defining measures, we talked about the measure $\mathbb{P}_X(B)$. In this section, we will construct tools to help us integrate over similar measures. First, we give the formal definition of measures of the form $\mathbb{P}_X(B)$.

DEFINITION 2.14 (Image measure). Suppose f is a measurable function from $(\Omega, \mathcal{F}, \mu)$ to $(\Omega', \mathcal{F}')$. The function $\mu_f : \mathcal{F}' \to [0, \infty]$ that sends an element A' of $\mathcal{F}$ to $\mu(f^{-1}(A'))$ is a measure on $(\Omega', \mathcal{F}')$ called the *image measure* of μ by f (also called the *pushforward* of μ). In the case of probability spaces, we call $\mathbb{P}_X$ the *probability distribution* of the random variable X.

Our main topic this subsection will be to understand how to work with integration with respect to image measures. First we discuss a somewhat easy result.

Proposition 2.15 (Transfer theorem)**.** *Let $f : (\Omega, \mathcal{F}, \mu) \to (\Omega', \mathcal{F}')$ be a measurable function. Suppose $\varphi : \Omega' \to \mathbb{R}$ is measurable with respect to the Borel algebra. Then $\varphi \in \mathcal{L}^1(\mu_f)$ if and only if $\varphi \circ f \in \mathcal{L}^1(\mu)$, in which case*

$$\int_{\Omega'} \varphi \, d\mu_f = \int_{\Omega} \varphi \circ f \, d\mu.$$

This also holds true without any integrability hypothesis if φ is positive.

Proof. Left as an exercise for the reader. The proof is the same as usual, start with indicator functions and build up to general functions. □

This proposition is quite general and we would like to specialize to a case where image measures have a very specific form.

DEFINITION 2.16 (Measures with density). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We will say a measure ν has a density g with respect to μ if

$$\nu(A) = \int_A g(\omega) \, d\mu(\omega)$$

.

It is with respect to those measure that integration will be most interesting for our purposes. In particular, we will be interested in changing measures on $\mathbb{R}^n$. This comes from the following change of variables formula which is quite handy. You might have seen this applied in vector calculus before, or differential geometry.

Theorem 2.17 (Change of variables)**.** *Suppose $f : \Omega \to \Omega'$ is a C^1-diffeomorphism between two open subsets of $\mathbb{R}^n$ and let $\rho : \Omega \to \mathbb{R}^+$ be a measurable function. Let μ be the measure of density ρ with respect to the usual (product) Lebesgue measure on Ω: $d\mu(x) = 1_\Omega(x)\rho(x)dx$. Then the image measure μ_f is $d\mu_f(y) = 1_{\Omega'}(y)\rho(f^{-1}(y))|\det Df^{-1}(y)|dy$. Thus, for any measurable function $\Phi : \Omega' \to \mathbb{R}^n$ that is either positive or in $\mathcal{L}^1(\mu_f)$, the following holds:*

$$\int_{\Omega} \Phi(f(x)) \, dm(x) = \int_{\Omega'} \Phi(y)\rho(f^{-1}(y))|\det Df^{-1}(y)| \, dy.$$

Proof. The proof can be found in any standard text on vector analysis. □

This summarizes the tools we will need to discuss probability properly. Before that however, we need to discuss spaces of functions and exactly what kind of convergence there can be between sequences of functions and their limits.

3 Function Spaces

3.1 Introduction

In basic real analysis, one learns early on about pointwise convergence. Usually, this is in the form of $\varepsilon - \delta$ proofs, where we construct a δ that depends on both x and ε. Then, later on, one realizes pointwise convergence is not strong enough to interchange limits between sequences and other objects and we require a stronger notion of convergence, namely that of uniform convergence. More specifically, this time we only allow δ to depend on ε and not at all on the x at hand. In a sense, the function itself converges to the limit, and not just each point at their own speed. In this section, we discuss the various relationships between convergences of functions and spaces of functions. We won't prove results in this chapter as they are all very standard results found in any text on functional analysis, and aren't the meat of probability theory.

3.2 $\mathbb{L}^p$ spaces

In the integration section, we have defined first the space of measurable functions, on which the integral is defined. Then, we've looked (for a measure μ) at the space $\mathcal{L}^1$. In this subsection, we want to expand this in two directions: First, we want to discuss whether there's a meaning to the space $\mathcal{L}^p$, for a general $p \geq 1$. After that, we wanna see if those spaces are *complete*, that is to say, whether Cauchy sequences always converge in the space itself.

Definition 3.1. Let $p \in [1, +\infty)$. We write $\mathcal{L}^p$ for the space of p-integrable functions, that is to say

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}} < +\infty$$

Notice that $\|\cdot\|_p$ is a norm on that space and thus $\mathcal{L}^p$ is a normed vector space.

Recall from functional analysis (or topology if you're Swiss) that every normed vector space has a Banach space completion. Effectively, this means there exists a complete space based on $\mathcal{L}^p$. Its existence is a standard exercise in functional analysis.

Theorem 3.2 (Completion)**.** *Let $\mathbb{L}^p$ denote the space of equivalence classes of functions in $\mathcal{L}^p$. We say that $f \sim g$ if and only if $f = g$ almost everywhere.*

Then $\mathbb{L}^p$ is the completion of $\mathcal{L}^p$. Furthermore, if $p = 2$, this is actually a Hilbert space with inner product defined as

$$\langle f|g\rangle_2 = \int_\Omega fg\, d\mu(\omega).$$

This theorem relies on a few fundamental inequalities that we state next.

Theorem 3.3. *For $f, g \in \mathbb{L}^p$, we have that:*

1. $\|f+g\|_p \leq \|f\|_p + \|g\|_p$ *(Minkowski's Inequality);*
2. $\|fg\|_1 \leq \|f\|_p \|g\|_q$ *where $\frac{1}{p} + \frac{1}{q} = 1$ (Hölder's Inequality).*

The case $p = 2$ in the second inequality is very important in its own right and is called the Cauchy-Schwarz *inequality. For now, we avoid talking about the case $p = 1$ where we have '$q = \infty$'.*

These new spaces give us a new way to talk about convergence. There is a subtle question of what it means exactly to integrate an equivalence class of functions, but do notice that for any two representatives of a class $[f]$, say f and f', their integral is equal. Indeed, this is exactly what we wanted and what we defined it to be. This is true because

$$\int_\Omega |f - f'|\, d\mu = \int_{N \cup N^c} |f - f'|\, d\mu = \int_{N^c} |f - f'|\, d\mu = 0$$

where N is a set of measure zero, and thus integrating over it gives us zero. Notice here, we implicitly assume that in this context, the product '$0 \cdot \infty' = 0$. This is a convention set to make sure everything holds. Going forward, we will just assume that by $f \in \mathbb{L}^p$, we mean a representative of the equivalence class of f. With that out of the way, let us define convergence in our new norm.

Definition 3.4. We say that f_n converges to f in $\mathbb{L}^p$, sometimes denoted by $f_n \xrightarrow{\mathbb{L}^p} f$, if

$$\lim_{n\to\infty} \|f_n - f\|_p = 0.$$

This gives us a new way to define convergence. In the next section, we finally introduce probability and we will look at even more modes of convergence.

4 Probability Theory

4.1 Introduction

In this section, we begin our study of probability by considering a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and real-valued random variables X, Y. Effectively, this means that $\mathbb{P}(\Omega) = 1$, and X, Y are measurable with respect to $\mathcal{F}$.

4.2 Expectation and Variance

In elementary probability theory, more specifically when we consider discrete random variables, we have the notion of expectation. Usually, this is defined as

$$\sum_{i=1}^{n} \mathbb{P}(X = x_i)x_i = \sum_{i=1}^{n} p_i x_i.$$

Using measure-theoretic tools, we will look to generalize this by treating this sum as a special form of a more general construct.

DEFINITION 4.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a real-valued random variable on it. Then the *expectation* of X is defined as

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, d\mathbb{P}(\omega).$$

This definition is rather abstract, and it is not so clear what the usual definition of expectation has to do with it. To relate it to our usual definition, we need to work with random variables that have a density. That is to say, those for which integrating against the probability measure gives us something that allows us to transfer the integral over to $\mathbb{R}$, which lets us apply calculus to it. More specifically:

DEFINITION 4.2 (Probability density). Let X be a real-valued random variable on Ω. Let $S \subseteq \mathbb{R}$. We say $f_X(x)$ is a *probability density* for X if

$$\begin{aligned}\mathbb{P}(X \in S) &= \mathbb{P}_X(S) \\ &= \int_{\mathbb{R}} 1_S f_X(x)\, dm(x) \\ &= \int_S f_X(x)\, dm(x).\end{aligned}$$

Such a function is necessarily positive a.e. and has integral equal to 1.

This is closely related but not the same as the following:

DEFINITION 4.3 (CDF). Let X be a real-valued random variable on Ω. The *cumulative distribution function* of X is defined as $F(t) = \mathbb{P}(X \leq t)$, for $t \in \mathbb{R}$. If X happens to have a density function f_X, then $\mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(x)\, dm(x)$. Furthermore, if F is differentiable at t, then $F'(t) = f_X(t)$.

We now apply the transfer theorem to understand how to integrate random variables with density.

Theorem 4.4 (Transferring probability measures)**.** *Suppose X has a density $f(x)$. This means that $\mathbb{P}_X(A) = 1_A(x)f(x)dm(x)$ where $dm(x)$ is the Lebesgue*

measure. Then for any measurable function $\Phi : \Omega \to \mathbb{R}^+$, we have that

$$\begin{aligned}\mathbb{E}\left[\Phi(X)\right] &= \int_\Omega \Phi \circ X \, d\mathbb{P} \\ &= \int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}_X(x) \\ &= \int_{\mathbb{R}} \Phi(x) f(x) \, dm(x)\end{aligned}$$

In particular, we have that

$$\mathbb{E}\left[X\right] = \int_{\mathbb{R}} x f(x) \, dm(x)$$

Proof. Apply the transfer theorem to X. □

Next we define the last two important concepts from probability theory we need to prove interesting theorems in the theory. The first one is fundamental, while we won't use the second one much but mention it for completion's sake.

DEFINITION 4.5 (Variance). Let X be a real-valued random variable. Then the *variance* of X denoted $\mathbb{V}(X)$ is defined as $\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]$. A quick calculation shows that $\mathbb{V}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$.

The final concept we wish to define in this subsection is that of *characteristic functions* (not to be confused with indicator functions). Those functions arise from applying Fourier analysis to probability theory.

Proposition 4.6. *Let X be a real-valued random variable on Ω. The* characteristic function *of X is defined as $\Phi_X(t) = \mathbb{E}\left[e^{itX}\right]$. The characteristic function of X completely determines the distribution of X. In particular, if X and Y are two random variables, then*

$$\Phi_X(t) = \Phi_Y(t) \iff F_X(t) = F_Y(t).$$

Proof. We don't concern ourselves with the proof here as the characteristic function won't be used anywhere else, we just mention it in passing. □

4.3 A Few Inequalities

In this subsection, we take a look at a few useful inequalities that will be necessary to prove some important theorems down the line. They are important in their own right however, which is why we discuss them now. Here, X always means a real-valued random variable.

Theorem 4.7 (Jensen's Inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. If X and $\varphi(X)$ are both integrable functions, then*

$$\varphi\left(\mathbb{E}\left[X\right]\right) \leq \mathbb{E}\left[\varphi(X)\right].$$

Proof. From convexity, we know that φ has at any point $x \in \mathbb{R}$ a left-derivative $\varphi'_g(x)$ and that

$$\varphi(y) - \varphi(x) \geq \varphi'_g(x)(y - x)$$

for any $y \in \mathbb{R}$. Thus, $\varphi(X) - \varphi(\mathbb{E}[X]) \geq \varphi'_g(\mathbb{E}[X])(X - \mathbb{E}[X])$. The result is obtained after taking the expectation of this inequality. □

Theorem 4.8 (Markov's Inequality)**.** *Let X be a real-valued random variable. Then, for all $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. It's enough to take the expectation of the inequality

$$t1_{X \geq t} \leq X1_{X \geq t} \leq |X|.$$

□

From this inequality, we get the following:

Theorem 4.9 (Bienaymé-Chebyshev's Inequality)**.** *If X^2 is integrable, then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{V}(X)}{t^2}.$$

Proof. Since $\{X \geq t\} \subset \{|X|^p \geq t^p\}$, we get from Markov's inequality that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[|X|^p]}{t^p}.$$

Take $Y = |X - \mathbb{E}[X]|$ and $p = 2$ and we get our result by applying the above to Y. □

4.4 Independence of Random Variables

In elementary probability theory, we have seen the notion of two events being independent. Usually, this is stated as

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

This sort of relation holds true and we will generalize it to σ-algebras first, and then to random variables themselves, thus answering the question: "What does it mean for two random variables to be independent?"

Definition 4.10 (Independence of σ-algebras). The events $\{A_i\}_{i \in I}$ are said to be *mutually independent* if for any finite family of indices $\{i_1, \ldots, i_n\}$ in I, we have that

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(A_{i_k})$$

Similarly, if $\{\mathcal{F}_i\}_{i\in I}$ is a family of σ-algebra on Ω, we say that they are *mutually independent* if for any finite family of indices $\{i_1, \ldots, i_n\}$in I, and for any choice of $A_{i_k} \in \mathcal{F}_{i_k}$ we have that

$$\mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}\right) = \prod_{k=1}^{n} \mathbb{P}(A_{i_k}).$$

This is consistent with what we've seen in elementary probability. Of interest to us is to generalize this notion to random variables themselves.

Definition 4.11 (Independence of random variables). Let $\{X_i\}_{i\in I}$ be a family of random variables. We say that they are *mutually independent* or more often simply *independent* if and only if the σ-algebras generated are (i.e., $\sigma(X_i)$ are independent.)

While this is a perfectly fine definition on its own, it's sometimes bothersome to check. Luckily, there's a nice proposition that summarizes equivalences between various ideas of independence for random variables. First, we need to discuss random vectors. Indeed, so far we've only talked about random variables from Ω to $\mathbb{R}$, but it makes perfect sense to think of random variables from Ω to $\mathbb{R}^n$ for some n. Those random variables will be called *random vectors*. They also have a probability distribution and a cumulative distribution function. Usually, we write $F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n)$ where $(t_1, \ldots, t_n)$ is a vector in $\mathbb{R}^n$. Again, some of those will have a density function. Henceforth, we will call it the *joint distribution density*. To contrast, the density function of a single random variable will be called the *marginal distribution density*. With that out of the way, here is the proposition.

Proposition 4.12. *The following are all equivalent:*

1. *The variables $\{X_i\}$ are mutually independent;*
2. *For any finite family of indices, the joint distribution of the random vector $(X_1, \ldots, X_n)$ is equal to the product of the distributions of each random variable, i.e. $\mathbb{P}_{X_{i_1},\ldots,X_{i_n}} = \prod_{k=1}^{n} \mathbb{P}_{X_{i_k}}$.*
3. *For any finite family of indices, and for any choice of measurable bounded functions $f_{i_k} : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}\left[f_{i_1}(X_1) \ldots f_{i_n}(X_n)\right] = \prod_{k=1}^{n} \mathbb{E}\left[f_{i_k}(X_k)\right]$.*

Proof. It is obvious that $2 \implies 3 \implies 1$. We deduce 2 from 1 as in the proof of Fubini's theorem, first for (f_{i_k}) indicator functions of measurable (Borel) sets, simple functions and finally, we apply the DCT. □

With this out of the way, we can now begin discussing the main modes of convergence that exist in probability theory.

5 Modes of Convergence

5.1 Introduction

We come now to the final section. In this section, we set out to do two things. The first one is to investigate what kind of convergence exists between sequences of random variables and their limits. The second one is to present two important theorems in probability and statistics: the law of large numbers and the central limit theorem.

5.2 Various notions of convergence

So far, we've encountered four different kinds of convergence. We will not be concerned too much with uniform convergence and pointwise convergence of random variables here. Instead, we will concern ourselves with convergence almost everywhere (that we will call convergence almost surely from now on), convergence in $\mathbb{L}^p$ and two new modes of convergence. We assume all our random variables are real-valued.

DEFINITION 5.1. Let (X_n) be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the sequence converges

1. **almost surely**, if
$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1;$$
2. in $\mathbb{L}^{\mathbf{p}}$, for $p \geq 1$, if X_n and X are in $\mathbb{L}^p$ and
$$\lim_{n\to\infty} \|f_n - f\|_p = 0;$$
3. **in probability**, if for all $\varepsilon > 0$, we have
$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0.$$
4. **in distribution** or **weakly**, if
$$\lim_{n\to\infty} F_n(x) = F(x),$$
for continuous points x, where F_n and F are the cumulative distribution functions of X_n and X respectively.

We denote convergence in probability with $X_n \xrightarrow{\mathbb{P}} X$, and convergence in distribution with $X_n \xrightarrow{distr.} X$. Convergence in distribution is the same as saying that the distribution and the characteristic functions of X_n converge to the distribution and characteristic function of X.

Notice this definition of convergence almost surely is precisely the same as converging except on possibly a set of measure zero. Convergence in probability is weaker however, as the following shows.

Proposition 5.2. *We have that $X_n \to X$ a.s. if and only if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{n\geq m} |X_n - X| \geq \varepsilon\right) = 0.$$

In particular, $X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{\mathbb{P}} X$.

Proof. By definition of the convergence of a sequence in $\mathbb{R}$ we get

$$\begin{aligned}\{X_n \to X\}^c &= \bigcup_{k\in\mathbb{N}^*} \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} \left\{|X_n - X| > \frac{1}{k}\right\} \\ &= \bigcup_{k\in\mathbb{N}^*} \bigcap_{m\in\mathbb{N}} \left\{\sup_{n\geq m} |X_n - X| \geq \frac{1}{k}\right\}.\end{aligned}$$

Thus, $X_n \to X$ a.s. if and only if

$$\forall k \in \mathbb{N}^* \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m\in\mathbb{N}} \sup_{n\geq m} |X_n - X| \geq \frac{1}{k}\right) = 0.$$

Using properties of measures discussed before, we know this to be equal to

$$\forall k \in \mathbb{N}^* \lim_{m\to\infty} \mathbb{P}\left(\sup_{n\geq m} |X_n - X| \geq \frac{1}{k}\right) = 0$$

But now we are done, since for any $\varepsilon > 0$ we can find a $k \in \mathbb{N}^*$ such that $\varepsilon > 1/k$ and this is equivalent to the result we were trying to prove. □

This is similar to the difference between uniform and pointwise convergence of real functions, where $f_n \xrightarrow{unif} f$ if and only if $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |f_n(x) - f(x)| = 0$. So we have now found one implication for our various modes of convergence. One could ask whether convergence in probability implies convergence almost surely. The answer in general is no, but there is a certain relationship nonetheless.

Proposition 5.3. *The sequence (X_n) converges to X in probability if and only if from any increasing sequence of natural numbers (n_k), there exists a subsequence (n_{k_j}) such that $X_{n_{k_j}} \xrightarrow{a.s.} X$.*

Proof. The proof relies on a few other results that we don't feel are necessary in the presentation of this material, so the proof is excluded. □

So while (X_n) itself might not converge almost surely to X, one of its subsequence does. Next we cover the case of convergence in $\mathbb{L}^p$.

Proposition 5.4. *Let $p \geq q \geq 0$. If $X_n \xrightarrow{\mathbb{L}^p} X$, then $X_n \xrightarrow{\mathbb{L}^q} X$. In particular, $\mathbb{L}^2$ convergence implies $\mathbb{L}^1$ convergence and $\mathbb{L}^1$ convergence implies convergence in probability.*

Proof. Let $p \geq q > 0$ and let $\alpha = p/q \geq 1$. From Jensen's inequality, we get

$$\mathbb{E}\left[|Y|^p\right] = \mathbb{E}\left[(|Y|^q)^\alpha\right] \geq \mathbb{E}\left[|Y|^q\right]^\alpha.$$

This shows $\mathbb{L}^p \subset \mathbb{L}^q$ and thus convergence in $\mathbb{L}^p$ implies convergence in $\mathbb{L}^q$. The fact this implies convergence in probability is related to the fact that convergence in probability is equivalent to convergence in $\mathbb{L}^0$ which we have not defined here. □

We now want to see if convergence in probability is equivalent to convergence in $\mathbb{L}^1$. Again, this isn't exactly true but if we can impose an additional requirement on our sequence, it holds.

Proposition 5.5 (Uniform integrability)**.** *A sequence of random variables is said to be* uniformly integrable *if*

$$\lim_{c\to\infty}\left(\sup_{n\in\mathbb{N}}\mathbb{E}\left[|X_n|1_{|X_n|\geq c}\right]\right) = 0.$$

Furthermore, if (X_n) is uniformly integrable, then it is bounded in $\mathbb{L}^1$. Conversely, if (X_n) is dominated by $Y \in \mathbb{L}^1$ or $(X_n) \subset \mathbb{L}^p$ for $p > 1$, then (X_n) is uniformly integrable.

Proof. We prove both parts separately.

1. We have that $\sup_n \mathbb{E}\left[|X_n|\right] \leq \sup_n \mathbb{E}\left[|X_n|1_{|X_n|\leq c}\right] + \sup_n \mathbb{E}\left[|X_n|1_{|X_n|\geq c}\right]$. The first term is bounded by c while the second is bounded because it is convergent.

2. Suppose $|X_n| \leq Y$. Then

$$\begin{aligned}\mathbb{E}\left[|X_n|1_{\{|X_n|\geq c\}}\right] &\leq \mathbb{E}\left[Y1_{\left\{Y\geq\sqrt{c}\right\}}\right] + \mathbb{E}\left[Y1_{\left\{Y\leq\sqrt{c}\right\}}1_{\{|X_n|\geq c\}}\right]\\ &\leq \mathbb{E}\left[Y1_{\left\{Y\geq\sqrt{c}\right\}}\right] + \sqrt{c}\mathbb{P}|X_n| \geq c\\ &\leq \mathbb{E}\left[Y1_{\left\{Y\geq\sqrt{c}\right\}}\right] + \frac{\sqrt{c}}{c}\mathbb{E}\left[|X_n|\right]\end{aligned}$$

   using Markov's inequality. The first term tends to 0 when $c \to \infty$ via the MCT. The second one is bounded by $\frac{\mathbb{E}[Y]}{c}$. Suppose now that (X_n) is bounded in $\mathbb{L}^p$ for p¿1. We get

$$\begin{aligned}\mathbb{E}\left[|X_n|1_{|X_n|\leq c}\right] &\leq \|X_n\|_p \mathbb{P}\left(|X_n| \geq c\right)^{\frac{1}{q}}\\ &\leq \|X_n\|_p \left(\frac{\mathbb{E}\left[|X_n|^p\right]}{c^p}\right)^{\frac{1}{q}}\end{aligned}$$

   by successively applying Hölder's and Markov's inequalities.

□

With this, we can state the theorem that relates $\mathbb{L}^1$ convergence and convergence in probability.

Theorem 5.6. *Let (X_n) be a uniformly integrable sequence of random variables. If (X_n) converges to X in probability, then (X_n) converges to X in $\mathbb{L}^1$.*

Proof. First, we need to settle the matter of integrability of X. From the characterization of convergence in probability, we know there exists a sequence X_{n_k} that converges to X almost surely. Using Fatou's lemma combined with the theorem above, we deduce

$$\mathbb{E}\left[|X|\right] \leq \liminf_{n_k \to \infty} \mathbb{E}\left[X_{n_k}\right] < \infty$$

which gives us integrability.
Let $Y_n = |X_n - X|$. Since (X_n) is uniformly integrable and X is integrable, Y_n is uniformy integrable (why?). Thus, for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}\left[Y_n\right] &= \mathbb{E}\left[Y_n 1_{Y_n \geq \varepsilon}\right] + \mathbb{E}\left[Y_n 1_{Y_n \leq \varepsilon}\right] \\ &\leq \mathbb{E}\left[Y_n 1_{Y_n \geq \varepsilon}\right] + \varepsilon.\end{aligned}$$

Choose $c > \varepsilon$ such that $\mathbb{E}\left[Y_n 1_{Y_n \geq c}\right] \leq \varepsilon$. Then

$$\begin{aligned}\mathbb{E}\left[Y_n 1_{Y_n \geq \varepsilon}\right] &\leq \mathbb{E}\left[Y_n 1_{Y_n \geq c}\right] + \mathbb{E}\left[Y_n 1_{c \geq Y_n \geq \varepsilon}\right] \\ &\leq \varepsilon + c\mathbb{P}\left(Y_n \geq \varepsilon\right).\end{aligned}$$

Thus,

$$\limsup_{n \to \infty} \mathbb{E}\left[Y_n\right] \leq 2\varepsilon$$

since $Y_n \to 0$ in probability. Since ε was arbitrary, we get the desired result. □

Lastly, we discuss the final implication we can get. We invite the reader to try and find counterexamples to the other implications.

Theorem 5.7. *If (X_n) converges to X in probability, then (X_n) converges to X in distribution. Convergence in distribution is thus the weakest form of convergence we have defined.*

Proof. The proof of this result relies on a technical lemma we do not need in the primary presentation of the material so we leave it out. □

Theorem 5.8. *The implications we've discussed are the only ones that exist between our various modes of convergence. For $q \geq p$, we have that*

$$\begin{gathered}(X_n) \xrightarrow{a.s.} X \implies (X_n) \xrightarrow{\mathbb{P}} X \implies (X_n) \xrightarrow{weakly} X, \\ (X_n) \xrightarrow{\mathbb{L}^q} X \implies (X_n) \xrightarrow{\mathbb{L}^p} X \implies (X_n) \xrightarrow{\mathbb{P}} X, \\ (X_n) \xrightarrow{\mathbb{P}+U.I.} X \iff (X_n) \xrightarrow{\mathbb{L}^1} X.\end{gathered}$$

Proof. Left to the reader as it gives good intuition in how modes of convergence work. The reader can consult Wikipedia or more academic sources for standard counterexamples. □

5.3 The Law of Large Numbers

We have developed much machinery to discuss probability. We are now able to define precisely what the word probability itself means, and how to work with its most basic objects. But how do we reconcile this with the more naive ideas of elementary probability? For instance, when we flip a fair coin, we know the probability of getting heads is exactly 1/2. But clearly, this doesn't mean that every time I flip a coin ten times in a row, I will get exactly 5 heads. How are we sure that our intuitive idea of probability is grounded in mathematics? This comes in the form of laws of large numbers. Briefly, the theorem tells us that the more we flip the coin, the closer the average value will get to the expectation. This is how casinos make money. Even if they get lucky winners from time to time, the law of large numbers provides a mathematical background that assures the casino that *in the long run*, it will make money. We only prove the weak law of large numbers and leave the proofs of the other two results to the mathematical literature.

DEFINITION 5.9. Let (X_n) be a sequence of random variables. The *sample average* up to n is the random variable

$$\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

It will be of interest to us to consider sequences of random variables that are both *independent and identically distributed*. This means that $F_{X_i} = F_{X_j}$ for all i, j in the index. Effectively, this means the joint distribution is the product of the marginal distributions, which are all equal. Furthermore, this implies that $\mathbb{E}[X_1] = \mathbb{E}[X_i]$ for all i. We will usually denote this expectation with μ.
Our theorem comes in two flavors, which we can now appreciate because of our work in the last section. As we have seen, convergence almost surely necessarily implies convergence in probability. This is the distinction between the two following theorems.

Theorem 5.10 (Weak Law of Large Numbers)**.** *Let (X_n) be a sequence of independent identically distributed random variables. Then the sample average converges in probability to the expectation. Symbolically:*

$$\overline{X_n} \xrightarrow{\mathbb{P}} \mu.$$

Proof. We assume that the variance of X_i is finite and equal to σ^2. The variance of $\overline{X_n}$ is equal to

$$\begin{aligned}\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{n\sigma^2}{n^2}\\ &= \frac{\sigma^2}{n}\end{aligned}$$

where the first equality comes from the independence of the random variables. Likewise, $\mathbb{E}\left[\overline{X_n}\right] = \mu$. Using Bienaymé-Chebyshev's inequality, we get

$$\mathbb{P}\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

From this, we get

$$\begin{aligned}\mathbb{P}\left(\left|\overline{X_n} - \mu\right| < \varepsilon\right) &= 1 - \mathbb{P}\left(\left|\overline{X_n} - \mu\right| \geq \varepsilon\right) \\ &= 1 - \frac{\sigma^2}{n\varepsilon^2}.\end{aligned}$$

Letting $n \to \infty$, we get $\overline{X_n} \xrightarrow{\mathbb{P}} \mu$ which proves the result. □

Theorem 5.11 (Strong Law of Large Numbers)**.** *If (X_n) is as above, then the sample average actually converges almost surely to the expectation. Symbolically:*

$$\overline{X_n} \xrightarrow{a.s.} \mu.$$

These tell us that in the long run, our usual interpretation of expectation as *the expected value* is correct. Likewise, we have a law of large numbers for our usual interpretation of probability itself.

Theorem 5.12 (Borel's Law of Large Numbers)**.** *Suppose we do repeated trials of a probabilistic experiment. Let E be an event and $p = \mathbb{P}(E)$ its probability. We let $N_n(E)$ denote the number of times E occurs in the first n trials. Then:*

$$\frac{N_n(E)}{n} \xrightarrow[n\to\infty]{a.s.} p$$

This is why we can expect to have approximately 50% of heads and 50% of tails in the long run after flipping fair coins for a long time. You can test this empirically by running a simulation.

5.4 The Central Limit Theorem

In our last subsection, we discuss an important theorem that has applications in statistics. We are again interested in understanding the asymptotic behavior of the sample average. We assume that all our random variables are independent and identically distributed (i.i.d.). Recall from elementary probability that the normal distribution for a random variable is of the form

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We know from the previous section that the sample average converges in probability and almost surely to the expectation. We are interested in understanding exactly how that happens. More specifically, we'd like to understand how the distribution itself changes as n tends to infinity. This is given by the following classical theorem.

Theorem 5.13 (Classical Central Limit Theorem)**.** *Suppose* $\mathbb{E}[X_i] = \mu$ *and* $\mathbb{V}(X_i) = \sigma^2 < \infty$*. Then, as* n *approaches infinity, the random variables* $\sqrt{n}\left(\overline{X_n} - \mu\right)$ *converge in distribution to a normal* $\mathcal{N}(0, \sigma^2)$*. Symbolically:*

$$\sqrt{n}\left(\overline{X_n} - \mu\right) \xrightarrow{distr.} \mathcal{N}(0, \sigma^2).$$

What does this imply for statistics? Well this explains why many density estimates have this bell curve shape. This comes from the shape of the normal distribution itself. If we apply this to the flipping coin example from last section, we will get that flipping many coins will give us a normal distribution for the number of heads (or tails, for that matter).

Appendices

A Common Distributions

In this appendix, we list a few common probability distributions for the reader. For the reader's sake, we also rewrite the change of variables formula in probabilistic term, so that they may use it to understand how products or sums modify the distributions of the random variables involved.

Theorem A.1 (Change of variables: probabilistic version)**.** *Suppose* X *is a real-valued random vector with density* f_X*. Then, if* ϕ *is a* C^1*-diffeomorphism, then the random variable* $Y = \phi(X)$ *has density*

$$g(y) = f\left(\phi^{-1}(y)\right)\left|\det D\phi^{-1}(y)\right|.$$

A.1 Discrete Random Variables

Example A.2 (Binomial distribution). Consider the following probabilistic experiment. You do n independent yes-no experiments, where yes has probability p and no has probability $q = 1 - p$. The *binomial distribution* is the probability distribution of a random variable X that counts the number of success in the sequence. We write $X \sim B(n, p)$ and we note that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

If $n = 1$, we call this the *Bernoulli distribution.*

Example A.3 (Poisson distribution). We say that X follows a *Poisson distribution* with parameter $\lambda > 0$ if

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

This distribution is usually used to compute the probability of a given number of events occurring in a fixed interval of time, provided these events occur at a constant mean rate and independently of the time since the last event.

A.2 Continuous Random Variables

EXAMPLE A.4 (Uniform distribution). The uniform distribution occurs when X has probability density function

$$f_X(x) = \frac{1}{b-a} 1_{[a,b]}(x).$$

EXAMPLE A.5 (Beta distribution). If X follows a Beta distribution (usually written $X \sim \text{Beta}(\alpha, \beta)$, then

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $B(\alpha, \beta)$ is a normalization constant equal to

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt.$$

EXAMPLE A.6 (Gamma distribution). We say that X follows a Gamma distribution ($X \sim \text{Gamma}(\alpha, \beta)$ if its density function is of the form

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} 1_{\{x \geq 0\}}$$

where $\Gamma(\alpha)$ is the Gamma function applied to α.

EXAMPLE A.7 (Exponential distribution). We say that X follows an exponential distribution ($X \sim \text{Exp}(\lambda)$) if

$$f_X(x) = \lambda e^{-\lambda x} 1_{\{x \geq 0\}}.$$

Furthermore, we have that $\lim_{n\to\infty} n\text{Beta}(1, n) = \text{Exp}(1)$.

EXAMPLE A.8 (Normal distribution). As we've seen before, a random variable has a *normal distribution* $\left(X \sim \mathcal{N}\left(\mu, \sigma^2\right)\right)$ if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

EXAMPLE A.9 (Cauchy distribution). This distribution usually has parameters but we will stick with the simplest version of it. We say X follows a *Cauchy distribution* if

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Try to compute the expectation and variance of this distribution.