THE TWO LIMIT THEOREMS OF PROBABILITY

SEBASTIEN VASEY

Contents

1. Introduction

The goal of these notes is to outline the proof of two fundamental theorems in probability. The setup of both theorems is the same: we have an infinite sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables with finite mean μ. We are interested in the behavior of their sum $S_n := \sum_{i=1}^{n} X_i$ as n becomes large. Using a frequentist interpretation of probability, one may expect that the average S_n/n of the random variables would become closer and closer to μ. This is indeed the case and the content of the *law of large numbers* (Theorem 4.4), the first theorem which we aim to prove.

More ambitiously, one may ask what precisely the distribution of S_n is. The answer is the content of the extraordinary *central limit theorem* (Theorem 4.5): if the X_i's have finite variance σ^2, then as n becomes large, S_n approaches a normal distribution with mean μn and variance $n\sigma^2$ (we will of course have to make precise what "approaches" exactly means).

2. Some terminology and notation

We have *not* defined what is meant by the mean of an arbitrary random variable (not necessarily discrete or continuous). Thus in these notes, by a *random variable* we mean a random variable that is either discrete or continuous (although one can make sense of the statements in general, see section 5.6 of Grimmett-Stirzaker). By a *function*, we mean a "reasonable" function that we can integrate (think "continuous, except perhaps at countably-many points").

We will also use the following bit of notation (5.6.A in Grimmett-Stirzaker):

Date: April 6, 2018.

Notation 2.1. Let X be a (discrete or continuous) random variable with distribution function F and let $g : \mathbb{R} \to \mathbb{C}$ be a function. We write $\int g dF$ or $\int g(x)dF$ for $\int_{-\infty}^{\infty} g(x)f_X(x)dx$ if X is continuous or $\sum_x g(x)f_X(x)$ if X is discrete.

Note in particular that $\int dF = 1$.

We have allowed complex-valued functions above (see the appendix for a crash course on complex numbers). We will also need to look at complex-valued random variables:

Definition 2.2. A *complex-valued random variable* is a function $Y : \Omega \to \mathbb{C}$ such that $\text{Re}(Y)$ and $\text{Im}(Y)$ are (real-valued) random variables.

We will always mention when we use complex-valued random variables. By default, "random variable" will mean a real-valued random variable.

Definition 2.3. Let Y be a complex-valued random variable. We define $\mathbb{E}(Y)$ by $\mathbb{E}(\text{Re}(Y)) + i\mathbb{E}(\text{Im}(Y))$.

One can check that the usual properties of expectation (for example: linearity, expectation of independent product is product of expectations) carry over to the case of complex-valued random variables.

2.1. **Continuity of integration.** When can one invert limits and integration? We will blackbox the following facts, whose proofs (in the continuous case) would need a precise definition of integration[1] (try to prove it for the discrete case!):

Fact 2.4 (Dominated convergence)**.** Let X be a random variable with distribution function F. Let $f_1, f_2 \ldots$ be a sequence of complex-valued functions which goes to f pointwise (that is, $f_n(x) \to f(x)$ for every x). If there exists a function g such that $|f_n(x)| \leq |g(x)|$ for all x and n and $\int |g|dF < \infty$, then $\int f_n dF$ goes to $\int f dF$.

Remark 2.5. A similar result also holds for limits of functions indexed by real numbers. Namely, let X be a random variable with distribution function F. Let f_t, $t \in \mathbb{R}$ be complex-valued functions and fix $a \in \mathbb{R}$. Assume that $f_t \to f$ as $t \to a$. If there exists a function g such that $|f_t(x)| \leq |g(x)|$ for all x and t and $\int |g|dF < \infty$, then $\int f_t dF$ goes to $\int f dF$ as $t \to a$. To see this from Fact 2.4, check that $\int f_{t_n} dF \to \int f dF$ for every subsequence (t_n) of real numbers going to a.

We deduce conditions under which one can invert the order of differentiatin and integration.

Fact 2.6. Let X be a random variable with distribution function F. Suppose that we are given $f : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$. Suppose that for all $x \in \mathbb{R}$, $\frac{d}{dt}f(t, x)$ exists (we will abuse notation write $f'(t, x)$ for $\frac{d}{dt}f(t, x)$). *If* there exists $g : \mathbb{R} \to \mathbb{C}$ such that for all $x, t \in \mathbb{R}$, $|f'(t, x)| \leq |g(x)|$ and $\int |g|dF < \infty$, *then* $\frac{d}{dt}\int f(t, x)dF = \int \frac{d}{dt}f(t, x)dF$.

Proof (optional). By definition of the derivative,

$$\frac{d}{dt}\int f(t, x)dF = \lim_{h\to 0}\int \frac{f(t + h, x) - f(t, x)}{h}dF$$

[1]The definition one can take for the purpose of these notes is called the *Lebesgue integral*. This is a generalization of the Riemann integral that one often studies in calculus classes. The latter is not powerful enough to satisfy the facts below, because a limit of Riemann-integrable function could fail to be Riemann-integrable (try to find an example!).

It is enough to see that we can put the limit inside the integration sign. Fix x, t, and $h > 0$. By the mean value theorem, there exists a point $c = c_x \in (t, t+h)$ such that $f'(c,x) = \frac{f(t+h,x)-f(t,x)}{h}$. Now, $|f'(c,x)| \leq |g(x)|$ by assumption. Thus $|\frac{f(t+h,x)-f(t,x)}{h}| \leq |g(x)|$, and so one can apply the dominated convergence theorem, as desired. □

The following result is also very useful. It says that under mild conditions one can invert the order of integration. We state it for integrals, and leave the corresponding statement for sums to the reader.

Fact 2.7 (The Fubini-Tonelli theorem)**.** Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$. Assume that at least one of the following conditions hold:

(1) $f(x,y) \geq 0$ (so it is a real number) for all x and y.
(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(x,y)| dxdy < \infty$.

Then:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dydx$$

Remark 2.8. To compute the integral in the second condition, one *can* change the order of integration, since the first condition holds for $|f|$.

Remark 2.9. Similar statements hold if the bounds are not $-\infty$ and ∞, but say $a \leq x \leq b$, $c \leq y \leq d$. One quick way to see this is to apply the result with the infinite bounds to $f \cdot \chi$, where χ is the indicator function of the set $[a,b] \times [c,d]$: $\chi(x,y)$ is 1 if $(x,y) \in [a,b] \times [c,d]$ and 0 otherwise.

3. Chebyshev's inequality

We first prove a special case of the law of large numbers using an important inequality.

Theorem 3.1 (Chebyshev's inequality)**.** Let X be a random variable and let $a > 0$. Then $P(|X| \geq a) \leq \frac{\mathbb{E}(X^2)}{a^2}$.

To see what Chebyshev's inequality says, consider the special case when X has mean zero and variance σ^2. Then we have that $P(|X| \geq a) \leq \frac{\sigma^2}{a^2}$. Thus X has to concentrate around its mean, in a way that is bounded by the variance. For example, setting $a = k\sigma$, we have that the probability that X is more than k standard deviations away from the mean is bounded by $\frac{1}{k^2}$. This is valid for *any* random variable, *regardless of its distribution!*

Since Chebyshev's inequality is very general, the bound it gives is often poor compared to specific cases[2]. For example, assume that $X \sim N(0,1)$. Then $P(|X| \geq k) = 2\frac{1}{\sqrt{2\pi}} \int_k^{\infty} e^{-x^2/2} dx$. We estimate:

$$\frac{2}{\sqrt{2\pi}} \int_k^{\infty} e^{-x^2/2} dx \leq \frac{\sqrt{2}}{\sqrt{\pi}} \int_k^{\infty} \frac{x}{k} e^{-x^2/2} dx = \frac{2}{\sqrt{\pi} k} e^{-k^2/2}$$

which is much smaller than $\frac{1}{k^2}$.

To prove Chebyshev's inequality, we first prove Markov's inequality:

[2]You will however show in your homework that the bound is sharp in general.

Theorem 3.2 (Markov's inequality)**.** Let X be a nonnegative random variable and let $a > 0$. Then $P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.

Proof. Let A be the event $\{X \geq a\}$. Then $X \geq aI_A$ (this uses that X is nonnegative), so taking expectations on both sides, $\mathbb{E}(X) \geq aP(X \geq a)$, and the result follows. □

Note that Markov's inequality can be used even when X has infinite variance (in which case Chebyshev cannot be used).

Proof of Chebyshev's inequality. Let $b := a^2$ and $Y := X^2$ and apply Theorem 3.2 to Y and b. We get that $P(X^2 \geq b) \leq \frac{\mathbb{E}(X^2)}{b}$. Taking square roots, we get that $P(|X| \geq \sqrt{b}) \leq \frac{\mathbb{E}(X^2)}{b}$, so $P(|X| \geq a) \leq \frac{\mathbb{E}(X^2)}{a^2}$, as desired. □

As a consequence of Chebyshev's inequality, let $X_1, X_2, \ldots$ be a sequence of independent and identically-distributed random variables with mean μ and variance σ^2. Let $S_n := \sum_{i=1}^n X_i$. We apply Chebyshev's inequality to $T_n := \frac{S_n}{n} - \mu$. We obtain that $P(|T_n| \geq a) \leq \frac{\mathbb{E}(T_n^2)}{a^2}$. Note that $\mathbb{E}(S_n) = n\mu$, hence $\mathbb{E}(T_n) = 0$. Thus $\mathbb{E}(T_n^2) = \mathrm{Var}(T_n)$. Further, $\mathrm{Var}(T_n) = \mathrm{Var}(\frac{S_n}{n}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$, using that the X_i's are independent. We obtain that $P(|T_n| \geq a) \leq \frac{\sigma^2}{a^2 n}$ which goes to zero as n goes to infinity. Thus $\frac{S_n}{n}$ concentrates around μ as n goes to infinity. To make this into a precise statement about the convergence of the random variable $\frac{S_n}{n}$, we need to talk about what convergence should mean.

4. Convergence of random variables

There are several senses in which a sequence $X_1, X_2, \ldots$ of random variables can converge to a random variable X (see Section 7.2 of Grimmett-Stirzaker for an overview). In these notes, we will only use one definition:

Definition 4.1. Let $X_1, X_2, \ldots$ be a sequence of random variables with distribution functions $F_{X_1}, F_{X_2}, \ldots$. Let X be a random variable with distribution function F_X. We say that (X_n) *converges to X in distribution*, written $X_n \xrightarrow{D} X$, if for every $x \in \mathbb{R}$ such that $P(X = x) = 0$, we have that $\lim_{n\to\infty} F_{X_n}(x) = F(x)$.

In other words, (X_n) converges to X in distribution if the sequence of distribution (F_{X_n}) converges pointwise to F_X *for each of the continuity points of* F_X. To see why we do *not* require that $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ when $P(X = x) \neq 0$, consider the following example: let X_n be the constantly $\frac{1}{n}$ random variable and let X be the constantly zero random variable. We would like to say that X_n goes to X. We have that $F_{X_n}(x) = 0$ if $x < 1/n$ and $F_{X_n}(x) = 1$ if $x \geq 1/n$, and $F_X(x) = 0$ if $x < 0$ and $F_X(x) = 1$ if $x \geq 0$. Observe that $\lim_{n\to\infty} F_{X_n}(0) = 0$, but $F_X(0) = 1$. Still, one can readily check that (X_n) converges to X in distribution, as $P(X = 0) \neq 0$. Intuitively, it does not matter if $F_{X_n}(x)$ does not converge to $F_X(x)$ when x is a discontinuity point: we will be able to recover what the random variable looks like in that neighborhood by adding a small epsilon to x. In fact, you will prove the following in your homework:

Exercise 4.2. Let X and Y be random variables[3]. Then:

[3]In fact, here we do not need that X and Y be either continuous or discrete.

(1) There are at most countably-many points x such that $P(X = x) > 0$.
(2) If $F_X(x) = F_Y(x)$ except for countably-many points x, then $F_X = F_Y$.

Hint: you may want to use the following two facts: a countable union of countable sets is countable, and any non-empty open interval of reals is not countable.

Convergence in distribution has several of the properties one would expect from a notion of convergence. You will explore this in your homework. For now, we define some more notation:

Definition 4.3. Let $X_1, X_2, \ldots$ be a sequence of random variable and let μ and σ be real numbers. We write $X_n \xrightarrow{D} \mu$ if $X_n \xrightarrow{D} X$, where X is the random variable that is constantly μ. We write $X_n \xrightarrow{D} N(\mu, \sigma^2)$ if $X_n \xrightarrow{D} X$, where $X \sim N(\mu, \sigma^2)$.

Using this notation, one can state the two limit theorems:

Theorem 4.4 (The law of large numbers)**.** Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite mean μ. Then:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{D} \mu$$

Theorem 4.5 (The central limit theorem)**.** Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite mean μ and variance σ^2. Then:

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}} \xrightarrow{D} N(0, \sigma^2)$$

Note that it is easy to check using linearity of expectation that the mean of $\frac{X_1+X_2+\ldots X_n-n\mu}{\sqrt{n}}$ is zero, and that its variance is σ^2 (using that the variance of a sum of independent random variables is the sum of the variance, and that $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$). From that point of view, the central limit theorem is at least plausible.

What cases of these theorems do we already know? We have already seen the *law of averages* (2.2 of Grimmett-Stirzaker) which is basically the law of large numbers when the X_i's are Bernoulli random variables. A harder argument (using Stirling's approximation formula) would yield the central limit theorem for Bernoulli random variables. We have also seen:

Exercise 4.6. If X and Y are independent $N(\mu, \sigma^2)$, $N(\lambda, \tau^2)$ random variables respectively, then $X + Y$ is $N(\mu + \lambda, \sigma^2 + \tau^2)$ and for a real number a, aX is $N(a\mu, a^2\sigma^2)$.

Thus if each $X_1, X_2, \ldots$ is a sequence of independent $N(\mu, \sigma^2)$ random variables, we have that $X_1 + \ldots + X_n$ is $N(n\mu, n\sigma^2)$, hence $\frac{X_1+\ldots+X_n}{n}$ is $N(\mu, \frac{\sigma^2}{n})$, and so it is easy to check that in distribution it will go to the constantly μ random variable. Similarly, $\frac{X_1+\ldots+X_n-n\mu}{\sqrt{n}}$ will be (exactly) $N(\mu, \sigma^2)$, as in the conclusion of the central limit theorem. Thus the central limit theorem holds for normal random variables.

In the rest of this section, we prove a special case of the law of large numbers using Chebyshev's inequality. The additional assumption compared to Theorem 4.4 is that the variables have finite variance.

Theorem 4.7 (The law of large numbers with finite variance)**.** Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite mean μ and variance σ^2. Then:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{D} \mu$$

Proof. We might as well assume that $\mu = 0$. If not, replace X_i by $Y_i = X_i - \mu$. Let $S_n = X_1 + \ldots + X_n$. By the argument at the end of Section 3, we have for each $a > 0$ that $P(|S_n/n| \geq a) \leq \frac{\sigma^2}{an}$. Let F_n be the distribution of S_n/n, and let F be the distribution of the constantly μ random variable. To establish convergence in distribution, we have to see that $F_n(x)$ goes to $F(x)$ whenever $x \neq \mu = 0$. Assume first that $x < 0$. Then $F(x) = 0$, and we have that $F_n(x) = P(S_n/n \leq x) \leq P(|S_n/n| \geq |x|) \leq \frac{\sigma^2}{xn}$, which goes to zero as n goes to infinity. Similarly, if $x > 0$, $F(x) = 1$ and $F_n(x) = P(S_n/n \leq x) = 1 - P(S_n/n > x) \geq 1 - P(|S_n/n| \geq x) \geq 1 - \frac{\sigma^2}{xn}$ which goes to 1 as n goes to infinity. □

5. Characteristic functions

Toward the proof of the general case of the law of large numbers and the central limit theorem, we have to develop tools to deal with sums of random variables. For certain discrete random variables, we saw the power of *generating functions*. Recall that the generating function of a discrete random variables X taking values in $\mathbb{N}$ was defined to be $G_X(s) = \mathbb{E}(s^X)$. We saw that generating functions of independent sums are product of the separate generating functions. For possibly continuous random variables it can be more convenient to work with the *moment generating function* $M_X(t) = \mathbb{E}(e^{tX})$. There are some issues with convergence, however[4], so we prefer to work with:

Definition 5.1. The *characteristic function* of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ defined by $\phi_X(t) = \mathbb{E}(e^{itX})$.

Remark 5.2. When X is a continuous random variable, $\phi_X(t) = \int_{-\infty}^{\infty} f_X(x) e^{itx} dx$. This is often called the *Fourier transform* of f_X.

Here, $\mathbb{C}$ is the set of *complex* numbers and $i = \sqrt{-1}$ (see the appendix for a crash course on complex numbers). Adding an i to the exponential makes a significant difference: $|e^{it}| = 1$ for any real number t. Thus the expectation always exists. In fact (see 5.7.3 in Grimmett-Stirzaker):

Lemma 5.3. Let X be a random variable and let ϕ be its characteristic function.

(1) $\phi(0) = 1$ and $|\phi(t)| \leq 1$ for all t.
(2) ϕ is uniformly continuous on $\mathbb{R}$.

Proof.

(1) $\phi(0) = \mathbb{E}(e^0) = \mathbb{E}(1) = 1$. Moreover:

$$|\phi(t)| = |\mathbb{E}(e^{ist})| = \left| \int e^{itx} dF_X \right| \leq \int |e^{itx}| dF_X = \int 1 dF_X = 1$$

[4]For example, the moment generating function of even an exponential random variable will not be defined everywhere.

(2) Fix t and h. We have to show that $|\phi(t+h) - \phi(t)|$ goes to zero as h goes to zero in a way that is independent from t. We compute:

$$|\phi(t+h) - \phi(t)| = |\mathbb{E}(e^{i(t+h)X} - e^{itX})| = |\mathbb{E}(e^{itX}(e^{ihX} - 1))| \leq \int |e^{ihx} - 1| dF_X$$

The right hand side does not depend on t and goes to zero as h goes to zero. Indeed by the triangle inequality, $|e^{ihx} - 1| \leq |e^{ihx}| + |1| = 2$, so the inside of the integral is bounded by the constant function 2, and $e^{ihx} - 1$ goes to zero as h goes to zero. By the dominated convergence theorem (Fact 2.4), $\int |e^{ihx} - 1| dF_X$ also goes to zero.

□

The characteristic functions has the following two basic properties, which turn out to be incredibly convenient: the characteristic function of an independent sum is the product of the individual characteristic functions, and the characteristic function of the product of a random variable by a constant also has a nice form:

Theorem 5.4.

(1) Let X and Y be independent random variables. Then $\phi_{X+Y} = \phi_X \phi_Y$.
(2) Let X be a random variable and let a be a real number. Then $\phi_{aX}(t) = \phi_X(at)$.

Proof.

(1) $\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX}e^{itY}) = \phi_X(t)\phi_Y(t)$, since X and Y are independent, and hence any function of X is also independent of any function of Y (4.2.3 in Grimmett-Stirzaker).
(2) Immediate.

□

Next, we investigate the meaning of derivatives of the characteristic function.

Lemma 5.5. Let k be a natural number, let X be a random variable and let ϕ be its characteristic function. If $\mathbb{E}(|X|^k) < \infty$, then $\phi^{(k)}(0) = i^k \mathbb{E}(X^k)$.

Proof. Let $F := F_X$. We prove by induction on k that $\phi^{(k)}(t) = \int i^k x^k e^{itx} dF$. From this, it follows that $\phi^{(k)}(0) = \int i^k x^k dF = i^k \mathbb{E}(X^k)$, as desired.

When $k = 0$, the result is immediate. Assume now that $k = n + 1$. We know by the induction hypothesis that $\phi^{(n)}(t) = \int i^n x^n e^{itx} dF$. We want to compute the derivative of $\phi^{(n)}$. We have to check that we can differentiate inside the integral sign. By Fact 2.6, we have to see that $\frac{d}{dt} i^n x^n e^{itx} = i^{n+1} x^{n+1} e^{itx}$ is uniformly bounded by an integrable function of x. Indeed, $|i^{n+1} x^{n+1} e^{itx}| = |x^{n+1}|$, and $\int |x^{n+1}| dF = \mathbb{E}(|X|^{n+1}) < \infty$ by assumption (since $\mathbb{E}(|X|^k) < \infty$, also $\mathbb{E}(|X|^m) < \infty$ for any $m \leq k$; this is an exercise in homework 10). Thus we obtain that $\phi^{(k)}(t) = \int i^k x^k e^{itx} dF$, as desired. □

It is time to give some examples of characteristic functions.

Example 5.6.

(1) The characteristic function of the constant random variable $X = a$ is $\phi_X(t) = e^{iat}$.
(2) If $X \sim \text{Bern}(p)$, then $\phi_X(t) = pe^{it} + 1 - p$.

(3) If $X \sim \text{Bin}(n,p)$, then X is a sum of n independent $\text{Bern}(p)$ random variables, so by Theorem 5.4, $\phi_X(t) = (pe^{it} + (1-p))^n$.
(4) If $X \sim N(\mu, \sigma^2)$, first set $Y = \frac{X-\mu}{\sigma}$. We have that $X = \sigma Y + \mu$ and $Y \sim N(0,1)$. By Theorem 5.4 and using the characteristic function of a constant computed above, $\phi_X(t) = \phi_Y(\sigma t)e^{i\mu t}$. It remains to compute ϕ_Y. Expanding the definitions, we have that $\phi_Y(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - x^2/2} dx$. To evaluate this integral, we use the following trick. As in the proof of Lemma 5.5, we can differentiate under the integral sign to obtain:

$$\phi'_Y(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ixe^{itx - x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ie^{itx} xe^{-x^2/2} dx$$

Integrating by parts, we obtain:

$$\phi'_Y(t) = \frac{1}{\sqrt{2\pi}} \left(-e^{-x^2/2} ie^{itx} \Big|_{x=-\infty}^{x=+\infty} - \int_{-\infty}^{\infty} te^{itx} e^{-x^2/2} \right) = -t\phi_Y(t)$$

To solve this differential equation, observe that $\frac{d}{dt} \log(\phi_Y(t)) = \frac{\phi'(t)}{\phi(t)} = -t$, hence integrating both sides, $\log(\phi_Y(t)) = \frac{-t^2}{2} + C$, for a constant C. We also know that $\phi_Y(0) = 1$, so $\log(\phi_Y(0)) = 0$, hence $C = 0$. Thus $\log(\phi_Y(t)) = \frac{-t^2}{2}$, so $\phi_Y(t) = e^{-t^2/2}$. This shows that in some sense the density function of a normal distribution is an eigenfunction for the Fourier transform and helps justify why the central limit theorem is true.

Just like for generating functions, the characteristic function of a random variable determines the distribution of the random variable. This is a fact from analysis which is not easy to prove: *if* ϕ_X is absolutely integrable[5] (i.e. $\int_{-\infty}^{\infty} |\phi(t)| dt < \infty$), then it turns out that X is a continuous random variable and we have the Fourier inversion formula:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) dt$$

This is still not easy to prove. In general, the integral on the right hand side may not converge. For example, if X is a constantly a "random" variable, then we have seen that $\phi_X(t) = e^{iat}$, so we end up having to compute an integral of the form $\int_{-\infty}^{\infty} e^{it(x-a)} dt = \int_{-\infty}^{\infty} \cos(t(x-a)) dt + i \int_{-\infty}^{\infty} \sin(t(x-a)) dx$, which does not converge. One way out of these difficulties is to slightly perturb what is inside the integral so that it does end up converging, and to approximate discrete random variables by continuous ones (for example the constantly a random variable can be approximated by a continuous random variable with a density function which strongly concentrates around a). Another way is to use improper integrals, interpreting $\int_{-\infty}^{\infty}$ by $\lim_{T \to \infty} \int_{-T}^{T}$. We obtain statements such as:

Fact 5.7. Let X be a random variable and let $a < b$ be such that $P(X = a) = P(X = b) = 0$. Then:

[5]This may not happen, for example if X is an exponential random variable with parameter λ, it turns out that the characteristic function is $\frac{\lambda}{\lambda - it}$, which is not absolutely integrable.

$$P(a \leq X \leq b) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) dt$$

The proof of Fact 5.7, while not out of reach, is a long and technical calculation that has little to do with probability (a long part of the proof revolves around computing the integral of $\frac{\sin(t)}{t}$), so we skip it and take Fact 5.7 as a black box. If you are interested, check out section 11.1 of Rosenthal's book (the full reference is in the syllabus).

As in the definition of convergence in distribution, it turns out that the condition that $P(X = a) = P(X = b) = 0$ does not matter too much: if this fails one can slightly change a and b so that the condition holds. Thus Fact 5.7 allows one to recover the distribution from the characteristic function:

Theorem 5.8 (Characteristic functions determine the distribution)**.** Let X and Y be random variables. If $\phi_X = \phi_Y$, then $F_X = F_Y$.

Proof. Let A be the set of real numbers x such that $P(X = x) > 0$ or $P(Y = x) > 0$. By Exercise 4.2, A is the union of two countable sets, hence countable. Fix $x \notin A$ and fix $\epsilon > 0$. We show that $|F_X(x) - F_Y(x)| < \epsilon$. Since ϵ is arbitrary, this will show that $F_X(x) = F_Y(x)$ and hence by Exercise 4.2 that $F_X = F_Y$. Since $x \notin A$, $P(X = x) = 0$. Now, we know that $\lim_{y \to -\infty} F(y) = 0$ for any distribution F, so pick a small-enough such that $F_X(a)$ and $F_Y(a)$ are both strictly less than $\frac{\epsilon}{2}$. The interval $(-\infty, a)$ is uncountable, hence there must exist $a' \in (-\infty, a)$ such that $a' \notin A$. By Fact 5.7, $F_X(x) - F_X(a') = P(a' \leq X \leq x) = P(a' \leq Y \leq x) = F_Y(x) - F_Y(a')$. Thus $F_X(x) - F_Y(x) = F_X(a') - F_Y(a')$, and by the triangle inequality, $|F_X(a') - F_Y(a')| \leq |F_X(a')| + |F_Y(a')| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, as desired. □

Another powerful fact that we will need is:

Fact 5.9 (The continuity theorem)**.** Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be a random variable. The following are equivalent:

(1) $X_n \xrightarrow{D} X$ (recall Definition 4.1).
(2) $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}$.

Again, we will assume this fact as a black box (see Section 11.1 of Rosenthal's book for the proof). The idea is that passing to the characteristic function "smoothes out" the distribution (for example, the characteristic function is uniformly continuous even though the distribution may not even be continuous), hence convergence in distribution is just the same as pointwise convergence of the characteristic functions.

6. Proofs of the two limit theorems

We are almost ready to prove the limit theorems. We recall one more fact from analysis[6]:

[6]This is usually stated for real-valued functions. However the same statement holds for complex-valued functions: one can simply apply real-valued Taylor's theorem to the real and imaginary parts separately.

Fact 6.1 (Taylor's theorem)**.** Let $f : \mathbb{R} \to \mathbb{C}$ be a function which is differentiable n times at zero. Then $f(x) = \left(\sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k\right) + h(x)$, for some function h where $\lim_{x \to 0} \frac{h(x)}{x^k} = 0$ (intuitively, h is much smaller than x^k).

Let us now prove the law of large numbers:

Proof of the law of large numbers (Theorem 4.4). Let $S_n := X_1 + X_2 + \ldots X_n$. We want to show that $\frac{S_n}{n} \xrightarrow{D} \mu$. For this we will pass to characteristic functions and use the continuity theorem. Let ϕ be the characteristic function of each X_i. We have that the characteristic function of S_n is $\phi_{S_n}(t) = (\phi(t))^n$, and hence the characteristic function of $\frac{S_n}{n}$ is $\phi_n(t) := (\phi(t/n))^n$ (Theorem 5.4). On the other hand, the characteristic function of the constantly μ random variable is $\phi_\mu(t) = e^{it\mu}$. By the continuity theorem, it suffices to see that $\phi_n(t) \to \phi_\mu(t)$ for each real number t. By Taylor's theorem, we can write $\phi(t/n) = \phi(0) + \phi'(0)(t/n) + h(t/n)$, where $\frac{h(t/n)}{t/n}$ goes to zero as n goes to infinity. Now $\phi(0) = 1$ and $\phi'(0) = i\mu$ (Lemma 5.5). Thus $(\phi(t/n))^n = \left(1 + i\mu\frac{t}{n} + h(t/n)\right)^n$. As n goes to infinity, this goes to $e^{it\mu}$, as desired. □

Proof of the central limit theorem (Theorem 4.5). The method of proof is similar to the proof of the law of large numbers. First, normalize by writing $Y_i := X_i - \mu$. Let $S_n := Y_1 + Y_2 + \ldots + Y_n = X_1 + \ldots + X_n - n\mu$. We want to show that $\frac{S_n}{\sqrt{n}} \xrightarrow{D} N(0, \sigma^2)$. Let ϕ be the characteristic function of the Y_i's and let ϕ_n be the characteristic function of $\frac{S_n}{\sqrt{n}}$. By Theorem 5.4, $\phi_n(t) = \phi(t/\sqrt{n})^n$.

Again, we consider the Taylor expansion of ϕ_n around zero. $\phi_n(t) = \phi(0) + \phi'(0)t + \phi''(0)\frac{t^2}{2} + h(t)$, where $\lim_{t \to 0} \frac{h(t)}{t^2} = 0$. We have that that $\phi(0) = 1$, and $\phi'(0) = i\mathbb{E}(Y_1) = 0$ (Lemma 5.5). By Lemma 5.5 again, $\phi''(0) = -\mathbb{E}(Y_1^2) = -\sigma^2$. So we obtain:

$$\phi_n(t/\sqrt{n}) = 1 - \sigma^2 \frac{t^2}{2n} + h(t/n)$$

So $\phi_n(t/\sqrt{n})^n = \left(1 - \frac{\sigma^2 t^2}{2n} + h(t/n)\right)^n$. Similarly to the proof of the law of large numbers, this goes to $e^{-\sigma^2 t^2/2}$ as $n \to \infty$. This is the characteristic function of a $N(0, \sigma^2)$ random variable (Example 5.6(4)), so by the continuity theorem, we are done. □

Appendix A. A crash course on complex numbers

A *complex number* is a pair (a, b) of real numbers. Let $\mathbb{C}$ be the set of complex numbers. We define the following operations on complex numbers:

- $(a, b) + (c, d) = (a + b, c + d)$.
- $(a, b) \cdot (c, d) = (ac - bd, ad + bc)$.

We let i be the complex number $(0, 1)$ and identify each real number a with the pair $(a, 0)$. Then (a, b) can be written $a + b \cdot i$, or just $a + bi$. We will adopt this notation throughout.

This is the formal definition, but the idea is that we think of i as a number satisfying the identity $i^2 = -1$. That is, i is the square root of -1. Then a complex

number is just a number of the form $a + bi$ with a, b real numbers. Addition and multiplication of complex numbers is defined as expected.

Each complex number $z = a + bi$ has a *real part*, $\mathrm{Re}(z) = a$ and an *imaginary part*, $\mathrm{Im}(z) = b$. One can think of complex numbers as points in the plane: the x-coordinate is their real part and the y-coordinate their imaginary part. The *absolute value (or modulus)* of a complex number $z = a + bi$ is then its distance from the origin $|z| := \sqrt{a^2 + b^2}$. We have the *triangle inequality*: $|z_1 + z_2| \leq |z_1| + |z_2|$ for all complex numbers z_1 and z_2. It is also easy to check that $|z_1 z_2| = |z_1||z_2|$ for all complex numbers z_1 and z_2.

We define concepts such as limits of sequences of complex numbers and integrals of functions $f : \mathbb{R} \to \mathbb{C}$ linearly. In details, if $z_1, z_2, \ldots$ is a sequence of complex numbers, we let $\lim_{n\to\infty} z_n$ be the complex number z (if it exists) such that $\mathrm{Re}(z) = \lim_{n\to\infty} \mathrm{Re}(z_n)$ and $\mathrm{Im}(z) = \lim_{n\to\infty} \mathrm{Im}(z_n)$. Similarly, if $f : \mathbb{R} \to \mathbb{C}$, we let $\int_a^b f(x)dx$ be $\int_a^b \mathrm{Re}(f(x))dx + i\int_a^b \mathrm{Im}(f(x))dx$ (defining what is meant by the integral of a function f whose *domain* is $\mathbb{C}$ is a different story that we will not get into). The following important property (which follows from the triangle inequality) is frequently useful:

$$\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|dx$$

We can define e^z, for a complex number z by $\sum_{n=0}^{\infty} \frac{z^n}{n!}$. One can prove (using the root test) that this converges absolutely everywhere. Expanding everything into a Taylor series, one can then show that the following important identity holds (called *Euler's formula*). For any real number t:

$$e^{it} = \cos(t) + i\sin(t)$$

Graphically, a complex number $z = a + bi$ with $|z| = 1$ is determined by the angle t between (a, b) and the x-axis. We have that $a = \cos(t)$ and $b = \sin(t)$. Euler's formula then says that $e^{it} = a + bi$. Note that this means in particular that $|e^{it}| = 1$ for any real number t. As a cute special case of Euler's formula, we have that $e^{i\pi} + 1 = 0$ (sometimes called *Euler's identity*).