Notes on weak convergence and related topics

Shota Gugushvili

Mathematical Institute, Faculty of Science, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
E-mail address: shota.gugushvili@math.leidenuniv.nl

2010 *Mathematics Subject Classification.* 60-01

Key words and phrases. Central limit theorem, sequential compactness, tightness, weak convergence, weak law of large numbers

Abstract. These notes deal with weak convergence of probability measures on the real line. They are largely based on the lecture notes written by Peter Spreij to accompany the *Measure Theoretic Probability* course.

Contents

Preface

These notes deal with weak convergence of probability measures on the real line and related topics. They are to a large extent based on the lecture notes written by Peter Spreij to accompany the *Measure Theoretic Probability* course. Other sources we used are listed in the bibliography.

Shota Gugushvili

CHAPTER 1

Weak convergence

1.1. Generalities

We start with the definition of weak convergence of probability measures on $(\mathbb{R}, \mathcal{B})$, and that of a sequence of random variables.

DEFINITION 1. *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$. It is said that μ_n converges weakly to μ, and we then write $\mu_n \xrightarrow{w} \mu$, if $\mu_n(f) \to \mu(f)$ for all $f \in C_b(\mathbb{R})$. If $X, X_1, X_2, \ldots$ are random variables (possibly defined on different probability spaces) with distributions $\mu, \mu_1, \mu_2, \ldots$, then we say that X_n converges weakly to X, and write $X_n \rightsquigarrow X$, if it holds that $\mu_n \xrightarrow{w} \mu$.*

Another accepted notation for weak convergence of a sequence of random variables is $X_n \xrightarrow{d} X$, and one says that X_n converges to X in distribution.

Consider the following example that illustrates for a special case that there is some reasonableness in Definition 1.

EXAMPLE 2. Let $\{x_n\}$ be a convergent of real numbers sequence with $\lim_{n\to\infty} x_n = 0$. Then for every $f \in C_b(\mathbb{R})$ one has $f(x_n) \to f(0)$. Let μ_n be the Dirac measure concentrated on x_n and μ the Dirac measure concentrated in the origin. Since $\mu_n(f) = f(x_n)$, we see that $\mu_n \xrightarrow{w} \mu$.

As a further result, here is a statement that gives an appealing sufficient condition for weak convergence of a sequence of random variables, when the random variables involved admit densities.

THEOREM 3. *Consider distributions $\mu, \mu_1, \mu_2, \ldots$ having densities $f, f_1, f_2, \ldots$ w.r.t. Lebesgue measure λ on $(\mathbb{R}, \mathcal{B})$. Suppose that $f_n \to f$ λ-a.e. Then $\mu_n \xrightarrow{w} \mu$.*

PROOF. We apply Scheffé's lemma to conclude that $f_n \to f$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$. Let $g \in C_b(\mathbb{R})$. Since g is bounded, we also have $f_n g \to fg$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ and hence $\mu_n \xrightarrow{w} \mu$. □

One could naively think of another definition of convergence of probability measures, for instance by requiring that $\mu_n(B) \to \mu(B)$ for every $B \in \mathcal{B}$, which is the same as to require that th class of test functions f consists of indicators of Borel sets, or even by requiring that the integrals $\mu_n(f)$ converge to $\mu(f)$ for every bounded measurable function. It turns out that each of these requirements is too strong to get a useful convergence concept. One drawback of such a definition is demonstrated by the following example with Dirac measures.

EXAMPLE 4. Assume the same setup as in Example 2 and take for concreteness $x_n = 1/n$. Let $B = (-\infty, x]$ for some $x > 0$. Then for all $x_n < x$, we have $\mu_n(B) = 1_B(x_n) = 1$ and $\mu(B) = 1_B(0) = 1$, so that $\mu_n(B) \to \mu(B)$. For $x < 0$ we

get that $\mu_n(B) = \mu(B) = 0$, and thus $\mu_n(B) \to \mu(B)$. But for $B = (-\infty, 0]$ we have $\mu_n(B) = 0$ for all n, whereas $\mu(B) = 1$. Hence convergence of $\mu_n(B) \to \mu(B)$ does not hold for this last choice of B, although it is quite natural in this case to say that $\mu_n \to \mu$. For the future reference note the following: if F_n is the distribution function of μ_n and F that of μ, then we have seen that $F_n(x) \to F(x)$, for all $x \in \mathbb{R}$, except for $x = 0$.

1.2. Criteria for weak convergence

In this section we give several criteria for weak convergence. These are primarily useful in the proofs.

THEOREM 5. *The following are equivalent:*

(i) $\mu_n \xrightarrow{w} \mu$;

(ii) every subsequence $\{\mu_{n_j}\}$ *of* $\{\mu_n\}$ *has a further subsequence* $\{\mu_{n_{j_k}}\}$, *such that* $\mu_{n_{j_k}} \xrightarrow{w} \mu$ *as* $k \to \infty$.

PROOF. That (i) implies (ii) is obvious. We prove the reverse implication. Assume the convergence $\mu_n \xrightarrow{w} \mu$ fails. This means there exists a bounded continuous function f, a subsequence $\{n_j\}$ of $\{n\}$ and a constant $\varepsilon > 0$, such that

$$|\mu_{n_j}(f) - \mu(f)| \geq \varepsilon$$

for all j. But then

$$|\mu_{n_{j_k}}(f) - \mu(f)| \geq \varepsilon$$

for any subsequence $\{n_{j_k}\}$ of $\{n_j\}$ as well. Hence $\{\mu_{n_j}\}$ has no further subsequence converging weakly to μ, a contradiction. □

Recall that the boundary ∂E of a set $E \in \mathcal{B}$ is $\partial E = \overline{E} \setminus E^\circ$, where $\overline{E}$ is the closure and E° is the interior of E. The distance from a point x to a set E is

$$d(x, E) = \inf\{|x - y| : y \in E\}.$$

The δ-neighbourhood of E (here $\delta > 0$) is the set $E^\delta = \{x : d(x, E) < \delta\}$.

The following result is known as the portmanteau lemma.

THEOREM 6 (Portmanteau lemma). *Let* $\mu, \mu_1, \mu_2, \ldots$ *be probability measures on* $(\mathbb{R}, \mathcal{B})$. *The following statements are equivalent.*

(i) $\mu_n \xrightarrow{w} \mu$.

(ii) $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ *for all closed sets* F.

(iii) $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$ *for all open sets* G.

(iv) $\lim_{n\to\infty} \mu_n(E) = \mu(E)$ *for all sets* E *with* $\mu(\partial E) = 0$ *(all* μ*-continuity sets).*

PROOF. We start with (i)⇒(ii). Given $\epsilon > 0$, choose $m \in \mathbb{N}$, such that for $\delta = 1/m > 0$, $\mu(F^\delta) < \mu(F) + \varepsilon$. This is possible, because F is closed and hence $F^\delta \downarrow F$ as $m \to \infty$. Let

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

and define

$$f(x) = \varphi\left(\frac{1}{\delta} d(x, F)\right).$$

Note that f is continuous, nonnegative, is bounded by 1 on $\mathbb{R}$, equals 1 on F and vanishes outside F^δ. Therefore,

$$\mu_n(F) = \int_F f \mathrm{d}\mu_n \leq \int_{\mathbb{R}} f \mathrm{d}\mu_n,$$

and

$$\int_{\mathbb{R}} f \mathrm{d}\mu = \int_{F^\delta} f \mathrm{d}\mu \leq \mu(F^\delta).$$

We also have

$$\lim_{n\to\infty} \int_{\mathbb{R}} f \mathrm{d}\mu_n = \int_{\mathbb{R}} f \mathrm{d}\mu.$$

Combining the above,

$$\limsup_{n\to\infty} \mu_n(F) \leq \lim_{n\to\infty} \int_{\mathbb{R}} f \mathrm{d}\mu_n = \int_{\mathbb{R}} f \mathrm{d}\mu \leq \mu(F^\delta) < \mu(F) + \varepsilon.$$

Since ε is arbitrary, the implication follows.

(ii)$\Leftrightarrow$(iii) follows by a simple complementation argument.

(ii) and (iii) together imply (iv) by

$$\begin{aligned}\mu(\overline{E}) &\geq \limsup_{n\to\infty} \mu_n(\overline{E}) \geq \limsup_{n\to\infty} \mu_n(E) \\ &\geq \liminf_{n\to\infty} \mu_n(E) \geq \liminf_{n\to\infty} \mu_n(E^\circ) \geq \mu(E^\circ),\end{aligned}$$

because $\mu(\partial E) = 0$ implies that the extreme terms are equal, the inequalities are in fact equalities, and so $\lim_{n\to\infty} \mu_n(E) = \mu(E)$.

(iv)$\Rightarrow$(i) Let $\varepsilon > 0$, $g \in C_b(S)$ and choose two constants C_1, C_2, such that $C_1 < g < C_2$. Let $D = \{x \in \mathbb{R} : \mu(\{g = x\}) > 0\}$. So, D is the set of atoms of g and hence it is at most countable (if not, μ would have an infinite total mass). Let $C_1 = x_0 < \ldots < x_m = C_2$ be a finite set of points not in D, such that $\max\{x_k - x_{k-1} : k = 1, \ldots, m\} < \varepsilon$. Let $I_k = (x_{k-1}, x_k]$. The continuity of g implies that if y is a boundary point of a set

$$\{x : x_{k-1} < g(x) \leq x_k\},$$

then $g(y)$ is either x_{k-1} or x_k. Hence the sets in the above display are μ-continuity sets. We have

(1.1)
$$\sum_{k=1}^m x_{k-1}\mu_n(x : x_{k-1} < g(x) \leq x_k) \leq \int_{\mathbb{R}} g \mathrm{d}\mu_n \leq \sum_{k=1}^m x_k \mu_n(x : x_{k-1} < g(x) \leq x_k),$$

and likewise,

(1.2)
$$\sum_{k=1}^m x_{k-1}\mu(x : x_{k-1} < g(x) \leq x_k) \leq \int_{\mathbb{R}} g \mathrm{d}\mu \leq \sum_{k=1}^m x_k \mu(x : x_{k-1} < g(x) \leq x_k).$$

Now note that the extreme terms in (1.1) converge to the respective extreme terms in (1.2). The latter differ by at most ε. Hence both the limit superior and limit inferior of $\int_{\mathbb{R}} g \mathrm{d}\mu_n$ are within distance ε of $\int_{\mathbb{R}} g \mathrm{d}\mu$. Since ε is arbitrary, the result follows. This finishes the proof of the theorem. □

Part (iv) of the portmanteau lemma is quite illustrative for understanding the definition of the weak convergence and in what way it differs from the requirement

$\mu_n(B) \to \mu(B)$ for every set B in the case of another would-be definition of weak convergence (cf. Section 1.1).

1.3. Convergence of distribution functions

In this section we give an appealing characterisation of weak convergence (convergence in distribution) in terms of distribution functions, which makes the definition of weak convergence look less abstract.

DEFINITION 7. *We shall say that a sequence of distribution functions $\{F_n\}$ on $\mathbb{R}$ converges weakly to a limit distribution function F, and shall write $F_n \rightsquigarrow F$, if $F_n(x) \to F(x)$ for all $x \in C_F$, where C_F is the set of all those points, at which F is continuous.*

THEOREM 8. *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on the real line and denote by $F, F_1, F_2, \ldots$ the corresponding distribution functions. The following statements are equivalent:*

(i) $\mu_n \xrightarrow{w} \mu$;
(ii) $F_n \rightsquigarrow F$.

PROOF. Assume (i). If x is a continuity point of F, the set $(-\infty, x]$, the boundary of which is $\{x\}$, is a μ-continuity set. Hence

$$F_n(x) = \mu_n((-\infty, x]) \to \mu((-\infty, x]) = F(x)$$

by the portmanteau lemma and thus (ii) holds.

Conversely, let (ii) hold. Fix an arbitrary $0 < \varepsilon < 1$ and pick two continuity points a and b of F in such a way that $F(a) < \varepsilon$ and $F(b) > 1 - \varepsilon$. Next, given $f \in C(\mathbb{R})$, choose the continuity points x_i of F, such that $a = x_0 < x_1 < \ldots < x_k = b$ and $|f(x) - f(x_i)| < \varepsilon$ for $x_{i-1} \le x \le x_i$ (this is possible by the uniform continuity of f on $[a, b]$). Define

$$S = \sum_{i=1}^{k} f(x_i)[F(x_i) - F(x_{i-1})], \quad S_n = \sum_{i=1}^{k} f(x_i)[F_n(x_i) - F_n(x_{i-1})].$$

By assumption, $S_n \to S$ as $n \to \infty$. Let $M = \sup_{x \in \mathbb{R}} |f(x)|$. We have

$$\left| \int_{\mathbb{R}} f \mathrm{d}\mu - S \right| < (2M + 1)\varepsilon.$$

Likewise,

$$\begin{aligned} \left| \int_{\mathbb{R}} f \mathrm{d}\mu_n - S_n \right| &\le \varepsilon + MF_n(a) + M(1 - F_n(b)) \\ &\to \varepsilon + MF(a) + M(1 - F(b)) \\ &< (2M + 1)\varepsilon. \end{aligned}$$

As a result,

$$\begin{aligned} \limsup_{n \to \infty} \left| \int_{\mathbb{R}} f \mathrm{d}\mu_n - \int_{\mathbb{R}} f \mathrm{d}\mu \right| &\le \limsup_{n \to \infty} \left| \int_{\mathbb{R}} f \mathrm{d}\mu_n - S_n \right| \\ &\quad + \left| \int_{\mathbb{R}} f \mathrm{d}\mu - S \right| + \lim_{n \to \infty} |S_n - S| \\ &\le 2(2M + 1)\varepsilon. \end{aligned}$$

Since ε is arbitrary, the limit superior on the left-hand side of the first inequality is in fact zero and the result follows. □

As shown in the next result, when the limit distribution function F is continuous everywhere, i.e. $C_F = \mathbb{R}$, the convergence $F_n(t) \to F(t)$ is in fact uniform in $t \in \mathbb{R}$.

THEOREM 9. *Suppose $F_n \rightsquigarrow F$ and F is continuous. Then*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| = 0.$$

PROOF. Let $k \in \mathbb{N}$ be fixed. By continuity of F and the intermediate value theorem, there exist points $-\infty = x_0 < x_1 < \ldots < x_k = \infty$, such that $F(x_i) = i/k$. Therefore, for $x_{i-1} \le x \le x_i$,

$$F_n(x) - F(x) \le F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + 1/k,$$
$$F_n(x) - F(x) \ge F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - 1/k.$$

Thus

$$|F_n(x) - F(x)| \le \sup_{0\le i\le k} |F_n(x_i) - F(x_i)| + 1/k, \quad x \in \mathbb{R}.$$

For any $\varepsilon > 0$, choose k so large that $1/k \le \varepsilon/2$. Next note that with this k, by convergence of $F_n(x) \to F(x)$ at all $x \in \mathbb{R}$, the supremum $\sup_{0\le i\le k} |F_n(x_i) - F(x_i)|$ can be made arbitrarily small, in particular smaller than $\varepsilon/2$, by taking n is large enough. Conclude that $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \varepsilon$ for all n large enough. Since ε is arbitrary, the result follows. □

1.4. Sequential compactness

In the previous sections we studied several alternative characterisations of weak convergence. In this section we will take a more abstract stance and study a condition guaranteeing that a sequence of probabaility measures has at least one weakly convergent subsequence. We first introduce the notion of sequential compactness of a sequence of probability measures.

DEFINITION 10. *A sequence of probability measures $\{\mu_n\}$ on $(\mathbb{R}, \mathcal{B})$ is called sequentially compact, if every subsequence $\{\mu_{n_k}\}$ of $\{\mu_n\}$ has a further weakly convergent subsequence.*

A general answer to the question whether a sequence $\{\mu_n\}$ is sequentially compact or not is given by Prokhorov's theorem. In its proof we need one auxiliary result, known as Helly's theorem.

The Bolzano-Weierstraß theorem states that every bounded sequence of real numbers has a convergent subsequence. The theorem easily generalises to sequences in $\mathbb{R}^d$, but fails to hold for uniformly bounded sequences in general metric spaces. But if extra properties are imposed, there can still be an affirmative answer. Something like that happens in Helly's theorem. At this point it is convenient to introduce the notion of a defective distribution function. Such a function, F say, has values in $[0, 1]$, is right-continuous and increasing, but at least one of the two properties $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$ fails to hold. The measure μ corresponding to F on $(\mathbb{R}, \mathcal{B})$ will then be a subprobability measure, $\mu(\mathbb{R}) < 1$.

THEOREM 11 (Helly's theorem). *Let $\{F_n\}$ be a sequence of distribution functions. Then there exists a possibly defective distribution function F and a subsequence $\{F_{n_k}\}$, such that $F_{n_k}(x) \to F(x)$, for all $x \in C_F$.*

PROOF. The main ingredient of the proof is an infinite repetition of the Bolzano-Weierstraß theorem combined with the Cantor diagonalisation. First we restrict ourselves to working on $\mathbb{Q}$ instead of $\mathbb{R}$, and exploit the countability of $\mathbb{Q}$. Write $\mathbb{Q} = \{q_1, q_2, \ldots\}$ and consider restrictions of F_n to $\mathbb{Q}$. Then the sequence $\{F_n(q_1)\}$ is bounded and along some subsequence $\{n_k^1\}$ it has a limit, $\ell(q_1)$ say. Look then at the sequence $\{F_{n_k^1}(q_2)\}$. Again, along some subsequence of $\{n_k^1\}$, call it $\{n_k^2\}$, we have a limit, $\ell(q_2)$ say. Note that along the thinned subsequence, we still have $\lim_{k\to\infty} F_{n_k^2}(q_1) = \ell(q_1)$. Continue like this to construct a nested sequence of subsequences $\{n_k^j\}$ for which we have that $\lim_{k\to\infty} F_{n_k^j}(q_i) = \ell(q_i)$ holds for every $i \le j$. Define a diagonal sequence $\{n_k\}$ by $n_k = n_k^k$. For an arbitrary i, along this sequence one has $\lim_{k\to\infty} F_{n_k}(q_i) = \ell(q_i)$. In this way we have constructed a function $\ell : \mathbb{Q} \to [0,1]$, and by the monotonicity of $F_n(t)$ in t this function is increasing.

In the next step we extend this function to a function F on $\mathbb{R}$ that is right-continuous, and still increasing. We put

$$F(x) = \inf\{\ell(q) : q \in \mathbb{Q},\, q > x\}.$$

Obviously, F is an increasing function. It is also right-continuous: let $x \in \mathbb{R}$ and $\varepsilon > 0$. There is $q \in \mathbb{Q}$ with $q > x$ such that $\ell(q) < F(x) + \varepsilon$. Pick $y \in (x, q)$. Then $F(y) < \ell(q)$ and we have $F(y) - F(x) < \varepsilon$, which shows that F is right-continuous. However, $\lim_{x\to\infty} F(x) = 1$ or $\lim_{x\to-\infty} F(x) = 0$ do not necessarily hold true. Thus F is a possibly defective distribution function.

We now show that $F_{n_k}(x) \to F(x)$ if $x \in C_F$. Fix $x \in C_F$ and let $\varepsilon > 0$. Pick q as above. By left-continuity of F at x, there is $y < x$ such that $F(x) < F(y) + \varepsilon$. Take now $r \in (y, x) \cap \mathbb{Q}$. Then $F(y) \le \ell(r)$, and hence $F(x) < \ell(r) + \varepsilon$. So we have the inequalities

$$\ell(q) - \varepsilon < F(x) < \ell(r) + \varepsilon.$$

Then

$$\limsup_{k\to\infty} F_{n_k}(x) \le \lim_{k\to\infty} F_{n_k}(q) = \ell(q) < F(x) + \varepsilon,$$
$$\liminf_{k\to\infty} F_{n_k}(x) \ge \liminf_{k\to\infty} F_{n_k}(r) = \ell(r) > F(x) - \varepsilon.$$

Since ε is arbitrary, the result follows. □

Here is an example, for which the limit in Theorem 11 is not a true distribution function.

EXAMPLE 12. Let μ_n be the Dirac measure concentrated on $\{n\}$. Then its distribution function is given by $F_n(x) = 1_{[n,\infty)}(x)$ and hence $\lim_{n\to\infty} F_n(x) = 0$. Hence the limit function F in Theorem 11 has to be the zero function, which is clearly defective. One colloquially says that in the limit the probability mass escapes to infinity.

Translated in terms of probability laws, Helly's theorem says that every sequence of probability measures $\{\mu_n\}$ has a (weakly) convergent subsequence, but that the limit law in general is a subprobability measure only. We are now interested in finding a condition, that would guarantee that the limit is a bona fide probability measure. A possible path is to require that all probability measures involved have probability one on a fixed bounded set. That would prevent the

phenomenon described in Example 12. However, this is a too stringent assumption, because it rules out many useful distributions. Fortunately, a considerably weaker assumption suffices. For any probability measure μ on $(\mathbb{R}, \mathcal{B})$ it holds that $\lim_{M\to\infty} \mu([-M, M]) = 1$. The next condition, tightness, gives a uniform version of this.

DEFINITION 13. *A sequence of probability measures $\{\mu_n\}$ on $(\mathbb{R}, \mathcal{B})$ is called tight, if $\lim_{M\to\infty} \inf_n \mu_n([-M, M]) = 1$.*

REMARK 14. Note that a sequence $\{\mu_n\}$ is tight iff every tail sequence $\{\mu_n\}_{n\geq N}$ is tight. In order to show that a sequence is tight it is thus sufficient to show tightness from a certain suitably chosen index on.

THEOREM 15 (Prokhorov's theorem[1]). *A sequence $\{\mu_n\}$ of probability measures on $(\mathbb{R}, \mathcal{B})$ is tight if and only if it is sequentially compact.*

PROOF. Suppose $\{\mu_n\}$ is sequentially compact, but not tight. Then there exists $\varepsilon > 0$, such that for any $M > 0$ and all n, $\mu_n([-M, M]^c) > \varepsilon$. It follows that for any $j \in \mathbb{N}$ and $I_j = (-j, j)$, one can find an index n_j, such that $\mu_{n_j}(I_j^c) > \varepsilon$. Extract from the sequence $\{\mu_{n_j}\}$ a weakly convergent subsequence $\{\mu_{n_{j_k}}\}$, and denote its weak limit by μ. By the portmanteau lemma, for every fixed $j \in \mathbb{N}$,

$$\limsup_{k\to\infty} \mu_{n_{j_k}}(I_j^c) \leq \mu(I_j^c).$$

Letting $j \to \infty$, we see that the right-hand side converges to zero, while the left-hand side stays bounded by $\varepsilon > 0$ from below. This contradiction proves the first implication.

We now prove the second implication. Let F_n be the distribution function of μ_n. By Helly's theorem, there exists a subsequence $\{F_{n_j}\}$ of the sequence of distribution functions $\{F_n\}$, such that $F_{n_j} \rightsquigarrow F$ as $j \to \infty$, for some, possibly defective, distribution function F. We will show that in fact

$$\lim_{x\to\infty} F(x) = 1, \quad \lim_{x\to-\infty} F(x) = 0, \tag{1.3}$$

so that F is a proper distribution function. By tightness of $\{\mu_n\}$, for any constant $0 < \varepsilon < 1$ there exists a constant $M_\varepsilon > 0$, such that $F_n(M_\varepsilon) > 1 - \varepsilon$ for all $n \in \mathbb{N}$. Without loss of generality, M_ε can be taken to be a continuity point of F. Then

$$F(M_\varepsilon) = \lim_{j\to\infty} F_{n_j}(M_\varepsilon) \geq 1 - \varepsilon.$$

Since ε is arbitrary, the above display and monotonicity of F imply the first equality in (1.3). The second one can be proved in a similar manner. This completes the proof. □

Theorem 15 has a simple corollary.

COROLLARY 16. *If $\mu_n \xrightarrow{w} \mu$ for some probability measure μ, then the sequence $\{\mu_n\}$ is tight.*

We also remark that tightness of a sequence $\{\mu_n\}$ in general is not sufficient for its weak convergence. Here is a simple counterexample: let $\mu_n = N(0, 1)$ for n odd and $\mu_n = N(0, 2)$ for n even. Then $\{\mu_n\}$ is tight, but does not converge weakly.

[1]The name Prokhorov is alternatively spelled as Prohorov, but Prokhorov is the way it appears in the English translation of the original paper containing (a much more general version of) the theorem. See Prokhorov (1956).

1.5. Continuous mapping theorem

The continuous mapping theorem is a result asserting that if a sequence of random variables $\{X_n\}$ converges in a suitable sense to a random variable X, then for a continuous function g the transformed sequence $\{g(X_n)\}$ converges to $g(X)$. We will prove a slightly more general result, that allows g to be discontinuous on a negligible set. Such a refinement does not require much additional technical effort, while occasionally being useful.

THEOREM 17 (Continuous mapping theorem). *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be continuous at every point of a set C, such that $\mathbb{P}(X \in C) = 1$.*

(i) If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.
(ii) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.
(iii) If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.

PROOF. Part (i) is trivial.

We prove part (ii). Let F be an arbitrary closed set. We have $\{g(X_n) \in F\} = \{X_n \in g^{-1}(F)\}$. Trivially, $g^{-1}(F) \subset \overline{g^{-1}(F)}$. Take an arbitrary $x \in \overline{g^{-1}(F)}$. By definition, there exists a sequence $\{x_m\}$, such that $x_m \to x$ and $g(x_m) \in F$. If $x \in C$, then $g(x_m) \to g(x)$, and $g(x) \in F$, because F is closed. Otherwise $x \in C^c$. Hence $\overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c$. Then by the portmanteau lemma and the fact that $\mathbb{P}(X \in C) = 1$,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(g(X_n) \in F) &\le \limsup_{n\to\infty} \mathbb{P}(X_n \in \overline{g^{-1}(F)}) \\ &\le \mathbb{P}(X \in \overline{g^{-1}(F)}) \\ &\le \mathbb{P}(X \in g^{-1}(F)) + \mathbb{P}(X \in C^c) \\ &= \mathbb{P}(g(X) \in F).\end{aligned}$$

By another application of the portmanteau lemma we conclude that $g(X_n) \rightsquigarrow g(X)$.

We move to part (iii). Assume that $g(X_n) \xrightarrow{\mathbb{P}} g(X)$ fails. Then there exist $\varepsilon > 0$, $\delta > 0$ and a subsequence $\{n_j\}$ of $\{n\}$, such that

$$\mathbb{P}(|g(X_{n_j}) - g(X)| > \varepsilon) > \delta. \tag{1.4}$$

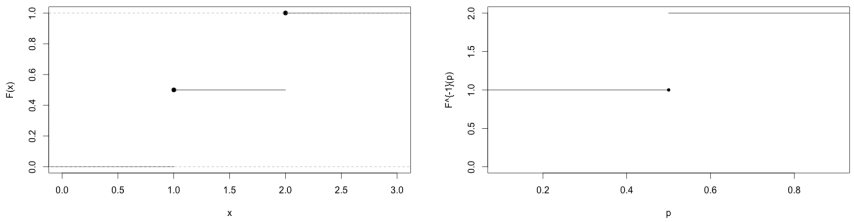
Extract from $\{n_j\}$ a further subsequence $\{n_{j_k}\}$, such that $X_{n_{j_k}} \xrightarrow{a.s.} X$. By part (iii), $g(X_{n_{j_k}}) \xrightarrow{a.s.} g(X)$. But this contradicts (1.4). The proof of the theorem is completed. □

It would be more appropriate, albeit clumsier, to call Theorem 17 the almost surely continuous mapping theorem.

EXAMPLE 18. Here is a simple illustration of Theorem 17. Let $Y_1, \dots, Y_n$ be an i.i.d. sample from the normal distribution with mean zero and unknown variance σ^2. By the strong law of large numbers,

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{a.s.} \sigma^2,$$

and hence $\hat{\sigma}^2$ is a reasonable estimator of σ^2. Since the function $g(x) = \sqrt{x}$ is continuous, $\hat{\sigma}_n$ is then a reasonable estimator of the standard deviation σ: we have $\hat{\sigma}_n \xrightarrow{a.s.} \sigma$.



(A) Distribution function (B) Quantile function

FIGURE 1. Distribution and quantile functions of the discrete uniform distribution on integers $1, 2$.

1.6. Almost sure representation theorem

Suppose we want to prove some distributional property of a sequence $\{X_n\}$ of random variables, knowing that $X_n \rightsquigarrow X$. In general this might be difficult, but is perhaps easier, if we knew that $X_n \xrightarrow{a.s.} X$. Unfortunately, the latter almost sure convergence might be difficult to establish, or perhaps is even false. However, the situation is not hopeless. The almost sure representation theorem, proved below, tells us that there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$, that supports random variables $\{\widetilde{X}_n\}, \widetilde{X}$, such that for all $n \in \mathbb{N}$, $\widetilde{X}_n \stackrel{d}{=} X_n$, $\widetilde{X} \stackrel{d}{=} X$, and $\widetilde{X}_n \xrightarrow{a.s.} \widetilde{X}$. We then prove the distributional property we are interested in for the sequence $\widetilde{X}_n$ on the space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. The result automatically carries over to the original sequence $\{X_n\}$.

We will need a number of results on quantile functions, which are of independent interest as well.

A distribution function in general is only non-decreasing, but not necessarily strictly increasing. Therefore, it typically does not admit the inverse function in the usual sense. Nevertheless, a kind of inverse, the quantile function, can still be defined. The quantile function of a distribution function F is a generalised inverse $F^{-1} : (0, 1) \mapsto \mathbb{R}$ given by

$$F^{-1}(p) = \inf \{x : F(x) \geq p\} .$$

For an illustration see Figure 1. The quantile function is left-continuous. Its range is equal to the support of F (or rather to the support of the corresponding probability measure μ; the support of a probability measure on $\mathbb{R}$ is defined as the set of all those points x, such that any open neighbourhood U_x of x has strictly positive measure: $\mu(U_x) > 0$. Intuitively, this is the smallest closed subset of $\mathbb{R}$ that receives measure 1 under μ (although you might be wondering at this point, this latter explanation is valid even for probability measures on separable metric spaces; see e.g. Theorem 2.1 on pp. 27–28 in Parthasarathy (2005)). As one example, the support of the standard normal distribution is the whole $\mathbb{R}$) and therefore, F^{-1} is often unbounded. An evident fact that the quantile function is monotone implies that it might have at most a countable number of discontinuity points only. The following lemma lists some other properties of F^{-1}. Of these we will only make partial use of (i)–(iv).

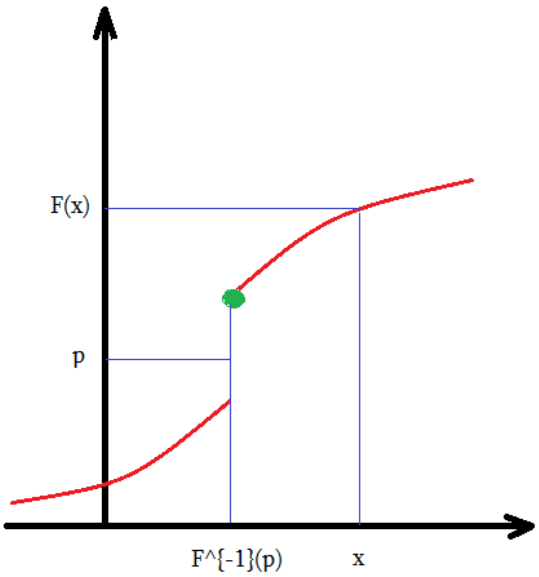


FIGURE 2. Distribution function (red line).

LEMMA 19. *For every* $0 < p < 1$ *and* $x \in \mathbb{R}$,

(i) $F^{-1}(p) \le x$ *if and only if* $p \le F(x)$;
(ii) $F \circ F^{-1}(p) \ge p$, *with equality holding if and only if* p *is in the range of* F; *the equality can fail if and only if* F *is discontinuous at* $F^{-1}(p)$;
(iii) $F_- \circ F^{-1}(p) \le p$, *with* $F_-(x) = F(x-)$;
(iv) $F^{-1} \circ F(x) \le x$; *the equality fails if and only if* x *is in the interior or at the right endpoint of a flat part of* F;
(v) $F \circ F^{-1} \circ F = F$; $F^{-1} \circ F \circ F^{-1} = F^{-1}$;
(vi) $(F \circ G)^{-1} = G^{-1} \circ F^{-1}$.

PROOF. (i) through (iv) can be proved either directly, by appealing to the definitions, or through a picture, such as the one given in Figure 2. To prove the first equality in (v), note that by (ii), the monotonicity of F and (iv),

$$F(x) = p \le F \circ F^{-1}(p) = F \circ F^{-1} \circ F(x) \le F(x).$$

The second equality in (v) follows from (ii), the monotonicity of F^{-1} and (iv) by $F^{-1}(q) \le F^{-1} \circ F \circ F^{-1}(q) \le F^{-1}(q)$. Finally, (vi) is a consequence of the definition of $(F \circ G)^{-1}$ and (i). □

As a consequence of (ii) and (iv), $F \circ F^{-1}(p) \equiv p$ and $F^{-1} \circ F(p) \equiv p$ on $(0,1)$ if and only if F is continuous and strictly increasing. In that case F^{-1} is a proper inverse of F, as it should be.

COROLLARY 20. *Let* F *be an arbitrary distribution function and* U *a uniform random variable on* $[0,1]$. *Then* $F^{-1}(U) \sim F$.

This follows from Lemma 19 (i). The transformation $F^{-1}(U)$ is called the quantile transformation.

COROLLARY 21. *Let $X \sim F$ for a continuous distribution function F. Then $F(X)$ is uniformly distributed on $[0,1]$.*

Again, this follows from Lemma 19 (i) and (ii) by

$$\begin{aligned}\mathbb{P}(F(X) \le x) &= \mathbb{P}(F(X) < x) = 1 - \mathbb{P}(F(X) \ge x) = 1 - \mathbb{P}(X \ge F^{-1}(x)) \\ &= \mathbb{P}(X < F^{-1}(x)) = \mathbb{P}(X \le F^{-1}(x)) = F \circ F^{-1}(x) = x,\end{aligned}$$

where $x \in (0,1)$. The transformation $F(X)$ for $X \sim F$ is called the probability integral transformation.

Quantile functions are occasionally useful when studying weak convergence of a sequence of random variables. In the next definition we introduce the notion of the weak convergence of a sequence of quantile functions.

DEFINITION 22. *We shall say that a sequence of quantile functions F_n^{-1} converges weakly to a limit quantile function F^{-1}, and denote this by $F_n^{-1} \rightsquigarrow F^{-1}$, if $F_n^{-1}(t) \to F^{-1}(t)$ at every point $0 < t < 1$, at which F^{-1} is continuous.*

Both the terminology and the notation for the weak convergence of quantile functions are reminiscent of those for the weak convergence of distribution functions. In fact, as shown in the next lemma, the two types of convergence are equivalent.

LEMMA 23. *For any sequence of distribution functions F_n, $F_n \rightsquigarrow F$ if and only if $F_n^{-1} \rightsquigarrow F^{-1}$.*

PROOF. Let U be a standard uniform random variable on some probability space, for instance on $([0,1], \mathcal{B}[0,1], \lambda)$. Since F^{-1} has at most a countable number of discontinuity points and the distribution of U is absolutely continuous, $F_n^{-1} \rightsquigarrow F^{-1}$ implies that $F_n^{-1}(U) \xrightarrow{a.s.} F^{-1}(U)$. Therefore, $F_n^{-1}(U) \rightsquigarrow F^{-1}(U)$. By Corollary 20, this is exactly the weak convergence $F_n \rightsquigarrow F$.

Now we prove the reverse implication. Let V be a standard normal random variable on some probability space, for instance on $([0,1], \mathcal{B}[0,1], \lambda)$, on which it can be obtained through the quantile transformation $\Phi^{-1}(U)$ for U a standard uniform random variable, see Corollary 20. Since the convergence $F_n(t) \to F(t)$ can fail only at a countable number of points t, and since the distribution of V is continuous, we have $F_n(V) \xrightarrow{a.s.} F(V)$ (and of course $F_n(V) \rightsquigarrow F(V)$). By Lemma 19 (i),

$$\begin{aligned}\Phi(F_n^{-1}(t)) &= \mathbb{P}(V < F_n^{-1}(t)) \\ &= 1 - \mathbb{P}(V \ge F_n^{-1}(t)) \\ &= 1 - \mathbb{P}(F_n(V) \ge t) \\ &= \mathbb{P}(F_n(V) < t),\end{aligned}$$

and similarly, $\mathbb{P}(F(V) < t) = \Phi(F^{-1}(t))$. By the portmanteau lemma,

$$\liminf_{n\to\infty} \mathbb{P}(F_n(V) < t) \ge \mathbb{P}(F(V) < t).$$

On the other hand, by elementary properties of the limits inferior and superior and the portmanteau lemma again,

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}(F_n(V) < t) &\le \limsup_{n\to\infty} \mathbb{P}(F_n(V) < t) \\ &\le \limsup_{n\to\infty} \mathbb{P}(F_n(V) \le t) \\ &= 1 - \liminf_{n\to\infty} \mathbb{P}(F_n(V) > t)\end{aligned}$$

$$\begin{aligned}&\leq 1-\mathbb{P}(F(V)>t)\\&=\mathbb{P}(F(V)\leq t).\end{aligned}$$

If $\mathbb{P}(F(V)\leq t)$ is continuous at t, then

$$\mathbb{P}(F(V)\leq t)=\mathbb{P}(F(V)<t)=\Phi(F^{-1}(t)),$$

and in this case

$$\begin{aligned}\liminf_{n\to\infty}\mathbb{P}(F_n(V)<t)&=\limsup_{n\to\infty}\mathbb{P}(F_n(V)<t)\\&=\lim_{n\to}\mathbb{P}(F_n(V)<t)\\&=\mathbb{P}(F(V)<t)\\&=\Phi(F^{-1}(t)).\end{aligned}$$

The function $\Phi(F^{-1}(\cdot))$ is certainly continuous at every point t, at which F^{-1} is. Since Φ^{-1} is a continuous function as well (cf. Lemma 19), from this it follows that $F_n^{-1}(t)\to F^{-1}(t)$ at every point t, at which F^{-1} is continuous. Thus $F_n^{-1}(t)\rightsquigarrow F^{-1}(t)$. □

The work we put in the previous results allows us to give a short proof of the almost sure representation theorem.

THEOREM 24 (Almost sure representation). *Let $X_n\rightsquigarrow X$. Then there exists a probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ and random variables $\widetilde{X}_n,\widetilde{X}$ defined on it, such that for all $n\geq 1$, $\widetilde{X}_n\overset{d}{=}X_n$, $\widetilde{X}\overset{d}{=}X$, and $\widetilde{X}_n\xrightarrow{a.s.}\widetilde{X}$.*

PROOF. Let F_n and F be the distribution functions of X_n and X, respectively. Consider the probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})=([0,1],\mathcal{B}[0,1],\lambda)$ and let U be a random variable on it with a standard uniform distribution. Define $\widetilde{X}_n=F_n^{-1}(U)$ and $\widetilde{X}=F^{-1}(U)$. By Corollary 20, $\widetilde{X}_n\overset{d}{=}X_n$ and $\widetilde{X}\overset{d}{=}X$. By Lemma 23, the convergence $F_n\rightsquigarrow F$ implies that $F_n^{-1}\rightsquigarrow F^{-1}$. By definition the latter means that $F_n^{-1}(t)\to F^{-1}(t)$ at all points t, at which F^{-1} is continuous. Note that F^{-1} has at most a countable number of discontinuity points, and hence the convergence $F_n^{-1}(t)\to F^{-1}(t)$ can perhaps fail only on a set with Lebesgue measure zero. Since U has a continuous distribution, this implies that $F_n^{-1}(U)\xrightarrow{a.s.}F^{-1}(U)$, i.e. $\widetilde{X}_n\xrightarrow{a.s.}\widetilde{X}$. □

Several applications of the almost sure representation theorem will be given in the next section.

1.7. Relation to other modes of convergence

Firstly, we show that convergence in probability implies convergence in distribution.

THEOREM 25. *Suppose that a sequence $\{X_n\}$ of random variables and a random variable X are defined on the same probability space. Assume that $X_n\xrightarrow{\mathbb{P}}X$. Then $X_n\rightsquigarrow X$.*

PROOF. Suppose the convergence $X_n\rightsquigarrow X$ fails. By definition this means that there exists $f\in C_b(\mathbb{R})$, such that the convergence $\mu_n(f)\to\mu(f)$ fails. Thus there exists $\varepsilon>0$ and a subsequence $\{n_k\}$ of $\{n\}$, such that $|\mu_{n_k}(f)-\mu(f)|\geq\varepsilon$ for all n_k. This is obviously true for any further subsequence of $\{n_k\}$ as well. Pick

a subsequence $\{n_{k_\ell}\}$ of $\{n_k\}$, such that $X_{n_{k_\ell}} \xrightarrow{a.s.} X$ (this is possible, because $X_n \xrightarrow{\mathbb{P}} X$). Then $\mu_{n_{k_\ell}}(f) \to \mu(f)$ by the dominated convergence theorem. But this leads to a contradiction that proves the theorem. □

COROLLARY 26. *Suppose that a sequence $\{X_n\}$ of random variables and a random variable X are defined on the same probability space. Assume that $X_n \xrightarrow{a.s.} X$. Then $X_n \rightsquigarrow X$.*

This follows from Theorem 25 and the fact that almost sure convergence implies convergence in probability.

The converse to Theorem 25 (and Corollary 26) is in general false.

EXAMPLE 27. Let $X \sim N(0,1)$ and $X_n = -X$ for all $n \in \mathbb{N}$. Then $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X| > \varepsilon/2) > 0$ for all $n \in \mathbb{N}$, and thus convergence in probability fails. Obviously, so does the almost sure convergence. On the other hand, by the symmetry of the standard normal distribution, $X_n \stackrel{d}{=} X$, and hence $X_n \rightsquigarrow X$.

There is one notable exception, however.

THEOREM 28. *Let the random variables $X, X_1, X_2, \ldots$ be defined on the same probability space. If $X_n \rightsquigarrow X$, where $\mathbb{P}(X = x) = 1$ for some $x \in \mathbb{R}$, then also $X_n \xrightarrow{\mathbb{P}} X$.*

PROOF. The distribution μ of X is the Dirac measure at x. For any $\varepsilon > 0$, the sets $(x + \varepsilon, \infty)$ and $(-\infty, x - \varepsilon)$ are μ-continuity sets. Note that

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n > x + \varepsilon) + \mathbb{P}(X_n < x - \varepsilon).$$

The right-hand side of the above display tends to zero as $n \to \infty$ by the portmanteau lemma. This completes the proof. □

Next we move to convergence of the first moments. Since weak convergence in general does not imply convergence in probability, neither does it in general imply convergence of means. But when the collection $\{X_n\}$ is uniformly integrable, the weak convergence $X_n \rightsquigarrow X$ can be strengthened to convergence of means: $\mathbb{E}[X_n] \to \mathbb{E}[X]$. The proof is a simple application of the almost sure representation theorem.

THEOREM 29. *Assume that $X_n \rightsquigarrow X$. If the sequence $\{X_n\}$ is uniformly integrable, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$.*

PROOF. By the almost sure representation theorem, there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ with random variables $\widetilde{X}, \widetilde{X}_1, \widetilde{X}_2 \ldots$, such that $X \stackrel{d}{=} \widetilde{X}$, $X_n \stackrel{d}{=} \widetilde{X}_n$ for all $n \in \mathbb{N}$, and $\widetilde{X}_n \xrightarrow{a.s.} \widetilde{X}$. By the uniform integrability of the family $\{X_n\}$, the family $\{\widetilde{X}_n\}$ is also uniformly integrable. Therefore $\mathbb{E}[\widetilde{X}_n] \to \mathbb{E}[\widetilde{X}]$, and since this latter convergence depends only on the laws of the random variables involved, the result follows. □

REMARK 30. Assume that $\{X_n\}$ and X are defined on the same probability space. Inspecting the proof of the previous theorem, one could have thought that not only do the means converge, but that we also have the $\mathcal{L}^1$-convergence: $\mathbb{E}[|X_n - X|] \to 0$. However, this in general is false and here is a simple counterexample: take $X \sim N(0,1)$ and $X_n = -X$ for all $n \in \mathbb{N}$. Then the conditions of Theorem 29 are satisfied, but $\mathbb{E}[|X_n - X|] = 2\mathbb{E}[|X|]$, which does not tend to zero. The point is that

$\mathbb{E}[|X_n - X|]$ depends on the bivariate law of (X_n, X), and this need not be the same as that of $(\widetilde{X}_n, \widetilde{X})$ (marginals do not determine joint distributions uniquely). This serves as a warning to when the almost sure representation theorem is applicable and when it is not: the representation does not in general preserve the dependence structure of $\{X_n\}$ and X, and hence typically cannot be used for statements dealing with multivariate vectors obtained from $\{X_n\}$ and X.

The following is what we can obtain without the uniform integrability assumption in Theorem 29. Again, the proof is an application of the almost sure representation theorem.

THEOREM 31. *If $X_n \rightsquigarrow X$, then $\liminf_{n\to\infty} \mathbb{E}[|X_n|] \geq \mathbb{E}[|X|]$.*

PROOF. By the almost sure representation theorem, there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ with random variables $\widetilde{X}, \widetilde{X}_1, \widetilde{X}_2 \ldots$, such that $X \stackrel{d}{=} \widetilde{X}$, $X_n \stackrel{d}{=} \widetilde{X}_n$ for all $n \in \mathbb{N}$, and $\widetilde{X}_n \xrightarrow{a.s.} \widetilde{X}$. Fatou's lemma implies that $\mathbb{E}[|\widetilde{X}|] \leq \liminf_{n\to\infty} \mathbb{E}[|\widetilde{X}_n|]$, and the statement follows. □

1.8. Slutsky's lemma

Suppose $X_n \rightsquigarrow X$ and the sequence $\{Y_n\}$ is close in some sense to $\{X_n\}$. What can be said about the weak limit of $\{Y_n\}$? Or suppose that $\{X_n\}$ and $\{Y_n\}$ are weakly convergent. What can be said about the weak convergence of the sequence $\{X_n Y_n\}$? The following result, known as Slutsky's lemma[2], gives an answer to these questions.

THEOREM 32. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of the random variables defined on the same probability space.*

(i) If $X_n \rightsquigarrow X$ and $|X_n - Y_n| \xrightarrow{\mathbb{P}} 0$, then $Y_n \rightsquigarrow X$.
(ii) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for a constant c, then $X_n Y_n \rightsquigarrow cX$.

PROOF. We first prove (i). Let F be closed and $\delta = 1/m$ for $m \in \mathbb{N}$. We have

$$\begin{aligned}\mathbb{P}(Y_n \in F) &= \mathbb{P}(X_n + Y_n - X_n \in F) \\ &= \mathbb{P}(X_n + Y_n - X_n \in F; |X_n - Y_n| < \delta) \\ &+ \mathbb{P}(X_n + Y_n - X_n \in F; |X_n - Y_n| \geq \delta) \\ &\leq \mathbb{P}(X_n \in \overline{F^\delta}) + \mathbb{P}(|X_n - Y_n| \geq \delta).\end{aligned}$$

Letting $n \to \infty$ and using the assumption $|X_n - Y_n| \xrightarrow{\mathbb{P}} 0$ and the portmanteau lemma, we obtain that

$$\limsup_{n\to\infty} \mathbb{P}(Y_n \in F) \leq \mathbb{P}(X \in \overline{F^\delta}).$$

Since $\overline{F^\delta} \downarrow F$ as $m \to \infty$, the result follows by another application of the portmanteau lemma.

Now we prove (ii). Write

$$X_n Y_n = X_n(Y_n - c) + cX_n. \tag{1.5}$$

[2]An alternative, but less common spelling of Slutsky's name is Slutzky. Also the result is at times called a theorem, not lemma.

An elementary argument shows that for any $\varepsilon > 0$ and $\delta > 0$,

$$\mathbb{P}(|X_n(Y_n - c)| > \varepsilon) \leq \mathbb{P}\left(|X_n| > \frac{\varepsilon}{\delta}\right) + \mathbb{P}(|Y_n - c| > \delta). \tag{1.6}$$

Fix ε and pick δ such that ε/δ and $-\varepsilon/\delta$ are continuity points of the distribution of X. Then the first term on the right-hand side of the above display converges to $\mathbb{P}(|X| > \varepsilon/\delta)$. The latter can be made arbitrarily small by taking δ small enough. As far as the second term in (1.6) is concerned, for every fixed δ it converges to zero. Hence $X_n(Y_n - c) \xrightarrow{\mathbb{P}} 0$. It is also easy to see that $cX_n \rightsquigarrow cX$ (this can be done in a variety of ways. For instance, the almost sure representation theorem and the dominated convergence theorem give for $f \in C_b(\mathbb{R})$ that $\mathbb{E}[f(cX_n)] \to \mathbb{E}[f(cX)]$). Now apply part (i) to the right-hand side of (1.5). □

Slutsky's lemma finds numerous applications in asymptotic theorems of mathematical statistics.

Exercises

1 Show that $X_n \rightsquigarrow X$ iff $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded uniformly continuous functions f.

2 Show the implication $F_n(x) \to F(x)$ for all $x \in C_F \Rightarrow \mu_n \xrightarrow{w} \mu$ without referring to the almost sure representation theorem. Hint: first you take for given $\varepsilon > 0$ a $K > 0$ such that $F(K) - F(-K) > 1 - \varepsilon$. Approximate a function $f \in C_b(\mathbb{R})$ on the interval $[-K, K]$ by a piecewise constant function, compute the integrals of this approximating function and use the convergence of $\{F_n\}$ at continuity points of F etc.

3 Let $\{\mu_n\}$ be a sequence of discrete uniform distributions on $[0, 1]$: $\mu_n(i/n) = 1/n$, $i = 1, \ldots, n$. Show that $\{\mu_n\}$ is weakly convergent and identify the weak limit.

4 Let $\{X_n\}$ be an i.i.d. sequence of exponentially distributed random variables: $F_{X_n}(x) = \mathbb{P}(X_n \leq x) = 1 - e^{-x}$ for $x \geq 0$ and $F_{X_n}(x) = 0$ for $x < 0$. Let $M_n = -\log n + \max_{1 \leq i \leq n} X_n$. Show that $F_{M_n} \rightsquigarrow F_M$, where $F_M(x) = \mathbb{P}(M \leq x) = e^{-e^{-x}}$, $x \in \mathbb{R}$. The latter distribution is known as the Gumbel distribution (or the extreme value distribution).

5 Consider the $N(\mu_n, \sigma_n^2)$ distributions, where the μ_n are real numbers and the σ_n^2 nonnegative. Show that this family is tight iff the sequences (μ_n) and (σ_n^2) are bounded. Under what condition do we have that the $N(\mu_n, \sigma_n^2)$ distributions converge to a (weak) limit? What is this limit?

6 Let random variabes X and X_n possess discrete distributions supported on $\mathbb{N}$. Show that $X_n \rightsquigarrow X$ if and only if $\mathbb{P}(X_n = m) \to \mathbb{P}(X = m)$ for every $m \in \mathbb{N}$.

7 Give an example of distribution functions F and F_n on the real line, such that $F_n \xrightarrow{w} F$, but $\sup_x |F_n(x) - F(x)| \to 0$ fails.

8 For a distribution function G on the real line the median is defined by $G^{-1}(1/2)$. Assume that $F_n \rightsquigarrow F$ and let $m = \text{med}(F)$ and $m_n = \text{med}(F_n)$ be the medians of F and F_n, respectively. Find suitable assumptions, under which $m_n \to m$ as $n \to \infty$.

9 Let F and G be two distribution functions on $\mathbb{R}$ and let

$$L(F, G) = \inf\{h > 0 : F(x - h) - h \leq G(x) \leq F(x + h) + h\}$$

be the Lévy distance between them (accept as a fact, or prove for yourself that $L(F, G)$ defines a distance). Show that the weak convergence $F_n \rightsquigarrow F$ is equivalent to convergence in the Lévy metric: $L(F_n, F) \to 0$. Hint: the implication $L(F_n, F) \to 0 \Rightarrow F_n \rightsquigarrow F$ follows from the definition. The other one can be established by contradiction.

10 Prove uniqueness of the weak limit μ of a weakly convergent sequence of probability measures μ_n.

CHAPTER 2

Characteristic functions

2.1. Definition and first properties

Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. X induces a probability measure on $(\mathbb{R}, \mathcal{B})$, the law or distribution of X, denoted by $\mathbb{P}^X$ or μ. This probability measure, in turn, determines the distribution function F of X. Conversely, F also determines $\mathbb{P}^X$. Hence distribution functions on $\mathbb{R}$ and probability measures on $(\mathbb{R}, \mathcal{B})$ are in bijective correspondence. In this chapter we develop another such correspondence. We start with a definition.

DEFINITION 33. *Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$. Its characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ is defined by*

$$\phi(u) = \int_{\mathbb{R}} e^{\imath u x} \mu(\mathrm{d}x). \tag{2.1}$$

Whenever needed, we write ϕ_μ instead of ϕ to express the dependence on μ.

Note that in this definition we integrate a complex valued function. By splitting a complex valued function $f = g + \imath h$ into its real part g and imaginary part h, we define $\int f \,\mathrm{d}\mu := \int g \,\mathrm{d}\mu + \imath \int h \,\mathrm{d}\mu$. For integrals of complex valued functions, previously shown theorems are, mutatis mutandis, true. For instance, one has $|\int f \,\mathrm{d}\mu| \leq \int |f| \,\mathrm{d}\mu$, where $|\cdot|$ denotes the norm of a complex number.

If X is a random variable with distribution μ, then ϕ_μ can alternatively be expressed as $\phi(u) = \mathbb{E}[\exp(\imath u X)]$. There are many random variables with distribution μ. They all have the same characteristic function. We also adopt the notation ϕ_X to indicate that we are dealing with the characteristic function of the random variable X.

Before we give some examples and elementary properties of characteristic functions, we look at a special case. Suppose that X admits a density f with respect to Lebesgue measure. Then

$$\phi_X(u) = \int_{\mathbb{R}} e^{\imath u x} f(x) \,\mathrm{d}x. \tag{2.2}$$

Analysts define for $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ the Fourier transform $\hat{f}$ by

$$\hat{f}(u) = \int_{\mathbb{R}} e^{-\imath u x} f(x) \,\mathrm{d}x.$$

What we thus see is the equality $\phi_X(u) = \hat{f}(-u)$. Given usefulness of Fourier transforms in various branches of mathematics, we then get a feeling that characteristic functions will be important in probability theory as well.

Computation of a characteristic function (if it is explicitly computable) is typically a clever exercise in integration.

EXAMPLE 34. Let $X \sim N(0,1)$. Then

$$\phi_X(u) = \mathbb{E}[e^{\imath uX}] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\imath ux} e^{-x^2/2} \mathrm{d}x = e^{-u^2/2}.$$

In fact,

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\imath ux} e^{-x^2/2} \mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \sum_{n=0}^{\infty} \frac{(\imath ux)^n}{n!} e^{-x^2/2} \mathrm{d}x \\
&= \sum_{n=0}^{\infty} \frac{(\imath u)^n}{n!} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^n e^{-x^2/2} \mathrm{d}x \\
&= \sum_{n=0}^{\infty} \frac{(\imath u)^n}{n!} \mathbb{E}[X^n].
\end{aligned}$$

For n odd, $\mathbb{E}[X^n] = 0$, while by Stein's lemma, see Lemma 35 ahead, for n even, $\mathbb{E}[X^n] = (n-1)!!$. Hence the above chain of equalities can be continued as

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{(\imath u)^n}{n!} \mathbb{E}[X^n] &= \sum_{n=0}^{\infty} \frac{(\imath u)^{2n}}{(2n)!} (2n-1)!! \\
&= \sum_{n=0}^{\infty} \frac{(\imath u)^{2n}}{(2n)!} \frac{(2n)!}{2^n n!} \\
&= \sum_{n=0}^{\infty} \left(-\frac{u^2}{2} \right)^n \frac{1}{n!} \\
&= e^{-u^2/2}.
\end{aligned}$$

Here we used the fact that

$$\begin{aligned}
(2n-1)!! &= \prod_{i=1}^{n} (2i-1) \\
&= \left\{ \prod_{i=1}^{n} (2i-1) \right\} \left\{ \prod_{i=1}^{n} (2i) \right\} \frac{1}{\prod_{i=1}^{n} (2i)} \\
&= \frac{(2n)!}{2^n n!}.
\end{aligned}$$

LEMMA 35 (Stein's lemma). *Let $X \sim N(0,1)$ and let g be a differentiable function satisfying $\mathbb{E}[g(X)X] < \infty$ and $\mathbb{E}[g'(X)] < \infty$. Then $\mathbb{E}[g(X)X] = \mathbb{E}[g'(X)]$.*

PROOF. We have

$$\mathbb{E}[g(X)X] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) x e^{-x^2/2} \mathrm{d}x.$$

By partial integration the right-hand side is equal to

$$-\frac{1}{\sqrt{2\pi}} g(x) e^{-x^2/2} |_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g'(x) e^{-x^2/2} \mathrm{d}x = \mathbb{E}[g'(X)].$$

This completes the proof. □

Here is another illustrative example.

EXAMPLE 36. Let X have a standard Cauchy distribution. Directly from the definition, when $u = 0$, $\phi_X(u) = 1$. Now assume $u \neq 0$. Then

$$\phi_X(u) = \frac{1}{\pi}\int_{\mathbb{R}} e^{\imath ux}\frac{1}{1+x^2}\mathrm{d}x = \frac{1}{\pi}\int_{\mathbb{R}}\frac{\cos(ux)}{1+x^2}\mathrm{d}x = |u|\frac{1}{\pi}\int_{\mathbb{R}}\frac{\cos(y)}{u^2+y^2}\mathrm{d}y.$$

The integral in the last equality is best evaluated through contour integration techniques. Let C_R be a closed contour consisting of the real line segment from $-R$ to R and the upper semi-circle Γ_R centred at the origin and of radius R. It can be shown that

$$\int_{\Gamma_R}\frac{e^{iz}}{u^2+z^2}\mathrm{d}z \to 0$$

as $R \to \infty$, see pp. 145–146 in Bak and Newman (2010). Therefore,

$$\int_{C_R}\frac{e^{iz}}{u^2+z^2}\mathrm{d}z \to \int_{\mathbb{R}}\frac{e^{iy}}{u^2+y^2}\mathrm{d}y.$$

Taking real parts on both sides, since $z_0 = \imath|u|$ is the only pole of the function $e^{iz}/(u^2+z^2)$ in the upper half plain, by the residue theorem we get that

$$\int_{\mathbb{R}}\frac{\cos(y)}{u^2+y^2}\mathrm{d}y = \mathrm{Re}\left[2\pi\imath\,\mathrm{Res}\left[\frac{e^{iz}}{u^2+z^2}, z_0\right]\right].$$

Now, since z_0 is a pole of order 2, it follows by (ii) on p. 130 in Bak and Newman (2010) that

$$\mathrm{Res}\left[\frac{e^{iz}}{u^2+z^2}, z_0\right] = \frac{\mathrm{d}}{\mathrm{d}z}\left[(z-z_0)^2\frac{e^{iz}}{u^2+z^2}\right]_{z=z_0} = \frac{e^{-|u|}}{2\imath|u|}.$$

Thus $\phi_X(u) = e^{-|u|}$ for $u \neq 0$. We conclude that $\phi_X(u) = e^{-|u|}$ for all $u \in \mathbb{R}$.

The following proposition lists some simple properties of characteristic functions.

PROPOSITION 37. *Let $\phi = \phi_X$ be the characteristic function of some random variable X. The following hold true:*

(i) $\phi(0) = 1$, $|\phi(u)| \leq 1$, *for all* $u \in \mathbb{R}$
(ii) ϕ *is uniformly continuous on* $\mathbb{R}$.
(iii) $\phi_{aX+b}(u) = \phi_X(au)e^{\imath ub}$.
(iv) ϕ *is real valued and symmetric around zero, if* X *and* $-X$ *have the same distribution.*
(v) If X *and* Y *are independent, then* $\phi_{X+Y}(u) = \phi_X(u)\phi_Y(u)$.
(vi) If $\mathbb{E}|X|^k < \infty$, *then* $\phi \in C^k(\mathbb{R})$ *and* $\phi^{(k)}(0) = \imath^k\mathbb{E}X^k$.

PROOF. Properties (i), (iii) and (iv) are trivial. Consider (ii). Fixing $u \in \mathbb{R}$, we consider $\phi(u+t) - \phi(u)$ for $t \to 0$. We have

$$\begin{aligned}|\phi(u+t) - \phi(u)| &= \left|\int(\exp(\imath(u+t)x) - \exp(\imath ux))\,\mu(\mathrm{d}x)\right| \\ &\leq \int|\exp(\imath tx) - 1|\,\mu(\mathrm{d}x).\end{aligned}$$

The functions $x \mapsto \exp(\imath tx) - 1$ converge to zero pointwise for $t \to 0$ and are bounded by 2. The result thus follows from dominated convergence.

Property (v) follows from the product rule for expectations of independent random variables.

Finally, property (vi) for $k = 1$ follows by an application of the dominated convergence theorem and the inequality $|e^{\imath x} - 1| \le |x|$, for $x \in \mathbb{R}$. The other cases can be treated similarly. □

REMARK 38. Here is a simple application of Proposition 37: if $X \sim N(m, \sigma)$, then $\phi_X(u) = e^{\imath um - \sigma^2 u^2/2}$.

REMARK 39. Warning: the converse to Proposition 37 (v) is typically false, i.e. from the equality

$$\phi_{X+Y}(u) = \phi_X(u)\phi_Y(u), \quad u \in \mathbb{R},$$

it cannot be concluded that X and Y are independent. Here is a counterexample: let X have a standard Cauchy distribution and let $Y = X$. Then

$$e^{-2|u|} = \phi_{2X}(u) = \phi_{X+Y}(u) = e^{-|u|}e^{-|u|} = \phi_X(u)\phi_Y(u),$$

although X and Y are obviously dependent in this case. More on this example later.

2.2. Inversion formula and uniqueness

Given a characteristic function ϕ, how can we find the corresponding distribution function F, or the corresponding law μ? As we will see, an answer to this qusetion is given by the inversion formula given below. Note that the integration interval in formula (2.3) is symmetric around zero. This is essential: an improper integral

$$\int_{-\infty}^{\infty} \frac{e^{-\mathrm{i}ua} - e^{-\imath ub}}{\imath u}\phi(u)\,\mathrm{d}u$$

typically does not exist (in the Lebesgue sense). That the limit in (2.3), called the Cauchy limit, is finite, is actually part of the assertion of the theorem.

THEOREM 40. *Let μ be a probability law and ϕ its characteristic function. Then for all $a < b$,*

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-\mathrm{i}ua} - e^{-\imath ub}}{\imath u}\phi(u)\,\mathrm{d}u = \mu((a,b)) + \frac{1}{2}\mu(\{a,b\}). \tag{2.3}$$

PROOF. We compute, using Fubini's theorem, which we will justify below,

$$\begin{aligned}
\Phi_T &:= \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-\imath ua} - e^{-\mathrm{i}ub}}{\imath u}\phi(u)\,\mathrm{d}u \qquad (2.4)\\
&= \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-\imath ua} - e^{-\imath ub}}{\mathrm{i}u}\int_{\mathbb{R}} e^{\imath ux}\,\mu(\mathrm{d}x)\,\mathrm{d}u\\
&= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{-T}^{T} \frac{e^{-\imath ua} - e^{-\mathrm{i}ub}}{\imath u}e^{\imath ux}\,\mathrm{d}u\,\mu(\mathrm{d}x)\\
&= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{-T}^{T} \frac{e^{\imath(x-a)u} - e^{\mathrm{i}(x-b)u}}{\imath u}\,\mathrm{d}u\,\mu(\mathrm{d}x) \qquad (2.5)\\
&=: \int_{\mathbb{R}} E_T(x)\,\mu(\mathrm{d}x)
\end{aligned}$$

Application of Fubini's theorem is justified as follows. First, the integrand in (2.5) is bounded by $b - a$, because $|e^{\imath x} - e^{\imath y}| \le |x - y|$ for all $x, y \in \mathbb{R}$. Second, the product measure $\lambda \times \mu$ on $[-T, T] \times \mathbb{R}$ is finite.

By splitting the integrand of $E_T(x)$ into its real and imaginary part, we see that the imaginary part vanishes and we are left with

$$\begin{aligned} E_T(x) &= \frac{1}{2\pi}\int_{-T}^{T} \frac{\sin(x-a)u - \sin(x-b)u}{u}\,\mathrm{d}u \\ &= \frac{1}{2\pi}\int_{-T}^{T} \frac{\sin(x-a)u}{u}\,\mathrm{d}u - \frac{1}{2\pi}\int_{-T}^{T} \frac{\sin(x-b)u}{u}\,\mathrm{d}u \\ &= \frac{1}{2\pi}\int_{-T(x-a)}^{T(x-a)} \frac{\sin v}{v}\,\mathrm{d}v - \frac{1}{2\pi}\int_{-T(x-b)}^{T(x-b)} \frac{\sin v}{v}\,\mathrm{d}v. \end{aligned}$$

The function g given by $g(s,t) = \int_s^t \frac{\sin y}{y}\,\mathrm{d}y$ is continuous in (s,t). Hence it is bounded on any compact subset of $\mathbb{R}^2$. Moreover, $g(s,t) \to \pi$ as $s \to -\infty$ and $t \to \infty$ (this can be shown by contour integration techniques; see e.g. pp. 146–147 in Bak and Newman (2010)[1]). Hence g, as a function on $\mathbb{R}^2$, is bounded in s,t. We conclude that also $E_T(x)$ is bounded as a function of T and x, the first ingredient to apply the dominated convergence theorem to (2.5), since μ is a finite measure. The second ingredient is to identify $E(x) := \lim_{T\to\infty} E_T(x)$. For an arbitrary α, a change of the integration variable $x = \alpha y$ gives

$$\int_0^\infty \frac{\sin(\alpha y)}{y}\,\mathrm{d}y = \operatorname{sgn}(\alpha)\frac{\pi}{2}.$$

Here $\operatorname{sgn}(\alpha)$ denotes $1, 0$ or -1 according to whether $\alpha > 0$, $\alpha = 0$ or $\alpha < 0$. By comparing the location of x relative to a and b, we use the value of the latter integral to obtain

$$E(x) = \begin{cases} 1 & \text{if } a < x < b, \\ \frac{1}{2} & \text{if } x = a \text{ or } x = b, \\ 0 & \text{else.} \end{cases}$$

We thus get, using the dominated convergence theorem again, that

$$\Phi_T \to \mu((a,b)) + \frac{1}{2}\mu(\{a,b\})$$

as $T \to \infty$. This completes the proof. □

REMARK 41. If a and b are continuity points of F, then the right-hand side of (2.3) is $F(b) - F(a)$. Thus ϕ determines F at all continuity points of F. But due to right-continuity of F, the latter completely determines F. F in turn determines μ, and so ϕ determines μ.

Let us now give another version of the inversion formula.

THEOREM 42. *If the characteristic function ϕ of a probability measure μ on $(\mathbb{R}, \mathcal{B})$ belongs to $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$, then μ admits a density f w.r.t. the Lebesgue measure λ. Moreover, f is continuous.*

PROOF. Define

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\imath u x}\phi(u)\,\mathrm{d}u. \tag{2.6}$$

[1] An alternative derivation is given here: `http://staff.science.uva.nl/~hvzanten/ex_5_9.pdf`

Since $|\phi|$ has a finite integral, f is well defined for every x. Observe that f is real valued, because $\overline{\phi(u)} = \phi(-u)$. An easy application of the dominated convergence theorem shows that f is continuous. Now note first that the limit of the integral in (2.3) is equal to the (Lebesgue) integral $\frac{1}{2\pi}\int \frac{e^{-\imath ua}-e^{-\imath ub}}{\imath u}\phi(u)\,\mathrm{d}u$, again because of dominated convergence. Next we use Fubini's theorem to compute for any continuity points $a < b$ of F that

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \frac{1}{2\pi}\int_a^b\int_{\mathbb{R}} e^{-\mathrm{i}ux}\phi(u)\,\mathrm{d}u\,\mathrm{d}x\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\phi(u)\int_a^b e^{-\mathrm{i}ux}\,\mathrm{d}x\,\mathrm{d}u\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\phi(u)\frac{e^{-\imath ua}-e^{-\mathrm{i}ub}}{\imath u}\,\mathrm{d}u\\ &= F(b)-F(a),\end{aligned}$$

where we also employed Theorem 40. Next, by continuity of $\int_a^b f(x)\,\mathrm{d}x$ in a and b, the relationship

$$\int_a^b f(x)\,\mathrm{d}x = F(b)-F(a)$$

in fact holds for any $a, b \in \mathbb{R}$. By continuity of f, for any $y \in [a,b]$ the Lebesgue integral $\int_a^y f(x)\,\mathrm{d}x$ equals the Riemann integral. By the fundamental theorem of calculus it follows that $F'(y) = f(y)$ for all $y \in (a,b)$ and so for all $y \in \mathbb{R}$. Since F is non-decreasing, f must be nonnegative, and hence it is a probability density. □

REMARK 43. Note the duality between the expressions (2.2) and (2.6). Apart from the presence of the minus sign in the integral and the factor 2π in the denominator in (2.6), the transformations $f \mapsto \phi$ and $\phi \mapsto f$ are similar.

The inversion theorem entails one very important result.

THEOREM 44. *Random variables X and Y are equal in distribution if and only if their characteristic functions are the same: $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$.*

PROOF. One side of the theorem is trivial. For the other side we argue as follows: suppose $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$. By Fubini's theorem and the inversion formula for characteristic functions, for every $\sigma_n > 0$ and $y \in \mathbb{R}$ we have

$$\begin{aligned}\int_{\mathbb{R}} e^{-\imath ty}e^{-\sigma_n^2 t^2/2}\phi_X(t)\mathrm{d}t &= \int_{\mathbb{R}} e^{-\imath ty}e^{-\sigma_n^2 t^2/2}\mathbb{E}[e^{itX}]\mathrm{d}t\\ &= \mathbb{E}\left[\int_{\mathbb{R}} e^{-\imath t(y-X)}e^{-\sigma_n^2 t^2/2}\mathrm{d}t\right]\\ &= \frac{\sqrt{2\pi}}{\sigma_n}\mathbb{E}\left[e^{-(y-X)^2/(2\sigma_n^2)}\right]\\ &= \frac{\sqrt{2\pi}}{\sigma_n}\int_{\mathbb{R}} e^{-(y-x)^2/(2\sigma_n^2)}\mathrm{d}F_X(x)\\ &= 2\pi f_{X+\sigma_n Z}(y).\end{aligned}$$

Here Z is a standard normal random variable independent of X and $f_{X+\sigma_n Z}$ is the density of $X + \sigma_n Z$ with respect to the Lebesgue measure. Replace ϕ_X with ϕ_Y in the above argument to see that $f_{X+\sigma_n Z}(y) = f_{Y+\sigma_n Z}(y)$. This implies that for

every $\sigma_n > 0$, $X + \sigma_n Z \stackrel{d}{=} Y + \sigma_n Z$. Letting $\sigma_n \to 0$ as $n \to \infty$, Slutsky's lemma gives that $X + \sigma_n Z \rightsquigarrow X$. Likewise, $X + \sigma_n Z \rightsquigarrow Y$. Due to the uniqueness of the weak limit, we then obtain that $X \stackrel{d}{=} Y$. □

Put another way, Theorem 44 implies that there is a one-to-one correspondence between probability measures and characteristic functions.

2.3. Necessary conditions

In the previous sections we have derived some properties a characteristic function possesses. Equally interesting is finding general conditions for a function ϕ to be a characteristic function. We will formulate two results in that direction. Their proofs can be found e.g. in Chung (2001) (see Theorems 6.5.2 and 6.5.3 there). The first result gives a necessary and sufficient condition, but is not easily verifiable. The second one is only sufficient, but its conditions are simpler.

Recall that a complex-valued function ϕ on $\mathbb{R}$ is called positive definite, if for any finite set of real numbers t_j's and complex numbers z_j's, $1 \leq j \leq n$, $n = 1, 2, \ldots$, we have

$$\sum_{j=1}^{n} \sum_{k=1}^{n} \phi(t_j - t_k) z_j \overline{z}_k \geq 0,$$

where $\overline{z}_k$ is a complex conjugate of z_k.

THEOREM 45 (Bochner-Khinchin theorem). *A function ϕ is a characteristic function if and only if it is positive definite, continuous at 0, and $\phi(0) = 1$.*

THEOREM 46 (Pólya's theorem). *Let ϕ satisfy the following conditions: $\phi(0) = 1$, ϕ is nonnegative, symmetric around zero, and decreasing, continuous and convex on $[0, \infty)$. Then ϕ is a characteristic function.*

EXAMPLE 47. Let $0 < \alpha \leq 1$. An application of Pólya's theorem gives that the function

$$\phi_\alpha(u) = e^{-|u|^\alpha}$$

is a characteristic function (check this). No such luck when $1 < \alpha < 2$, but via an alternative route ϕ_α can nevertheless be shown to be a characteristic function in that case as well (see e.g. pp. 192–193 in Chung (2001)). When $\alpha = 2$, we know that ϕ corresponds to the normal distribution. A probability distribution that has ϕ_α as a characteristic function is called a stable distribution with index α. We finally remark that it can be shown that ϕ_α with $\alpha > 2$ is not a characteristic function (in this case ϕ_α is twice differentiable at zero and $\phi'_\alpha(0) = \phi''_\alpha(0) = 0$. Assume ϕ_α is a characteristic function. By Theorem 6.4.1 in Chung (2001) the first and second moments of the corresponding probability law are zero. But then μ must be the Dirac measure at zero, so that $\phi_\alpha(u) = 1$ for all $u \in \mathbb{R}$. This is a contradiction).

2.4. Multidimensional case

Our treatment in this section is cursory and we omit most details.

The characteristic function ϕ of a probability measure μ on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ is defined by the k-dimensional analogue of (2.1). We have with $u, x \in \mathbb{R}^k$, $\langle \cdot, \cdot \rangle$ the standard inner product,

$$\phi(u) = \int_{\mathbb{R}^k} e^{\imath \langle u, x \rangle} \mu(\mathrm{d}x).$$

Like in the real case, also here probability measures are uniquely determined by their characteristic functions. As a consequence we have the following characterization of independent random variables.

PROPOSITION 48. *Let $X = (X_1, \dots, X_k)$ be a k-dimensional random vector. Then $X_1, \dots, X_k$ are independent random variables iff $\phi_X(u) = \prod_{i=1}^k \phi_{X_i}(u_i)$, $\forall u = (u_1, \dots, u_k) \in \mathbb{R}^k$.*

PROOF. If the X_i are independent, the statement about the characteristic functions is proved in the same way as Proposition 37 (v). If the characteristic function ϕ_X factorizes as stated, the result follows by the uniqueness property of characteristic functions. □

REMARK 49. Let $k = 2$ in the above proposition. If $X_1 = X_2$ as in Remark 39, then we do not have $\phi_X(u) = \phi_{X_1}(u_1)\phi_{X_2}(u_2)$ for every u_1, u_2 (you check!), in agreement with the fact that X_1 and X_2 are not independent. But for the special choice $u_1 = u_2$ this product relation holds true.

EXAMPLE 50. Let X and Y be independent standard normal random variables. Then somewhat unexpectedly, the random variables $X - Y$ and $X + Y$ are also independent, which can be shown using Proposition 48.

Exercises

1 Let ϕ be a characteristic function. Show that so is $|\phi|^2$.

2 If F and G are distribution functions, such that $F = \sum_{j=1}^m b_j \delta_{a_j}$ and G has a density, say g, show that the convolution $F * G$ has a density and find it.

3 Show that for any characteristic function ϕ,

$$\operatorname{Re}[1 - \phi(u)] \geq \frac{1}{4} \operatorname{Re}[1 - \phi(2u)].$$

4 A random variable X with the characteristic function ϕ is symmetric, if and only if $\phi(u)$ is real for all $u \in \mathbb{R}$.

5 Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables and N a Poisson(λ) distributed random variable, independent of the X_n. Put $Y = \sum_{n=1}^N X_n$. Let ϕ be the characteristic function of the X_n and ψ the characteristic function of Y. Show that $\psi = \exp(\lambda\phi - \lambda)$.

6 If X has an exponential distribution with parameter λ, then $\phi_X(u) = \lambda/(\lambda - \mathrm{i}u)$.

7 Let ϕ be a real characteristic function with the property that $\phi(nu) = \phi(u)^n$ for all $u \in \mathbb{R}$ and $n \in \mathbb{N}$. Show that for some $\alpha \geq 0$ it holds that $\phi(u) = \exp(-\alpha|u|)$. Let X have characteristic function $\phi(u) = \exp(-\alpha|u|)$. If $\alpha > 0$, show that X admits the density

$$x \mapsto \frac{\alpha}{\pi} \frac{1}{\alpha^2 + x^2}.$$

What is the distribution of X if $\alpha = 0$?

8 Prove the statement made in Example 50. Also verify that the function ϕ_α from Example 47 is indeed a characteristic function for $0 < \alpha \leq 1$.

9 Let μ be a probability law on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and let ϕ be the corresponding characteristic function. Show that for any fixed $x \in \mathbb{R}$,

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-\imath u x} \phi(u) \, \mathrm{d}u = \mu(\{x\}).$$

Hint: reduce the question to studying

$$\int_{\mathbb{R}\setminus\{x\}} \frac{\sin(T(y-x))}{T(y-x)}\mu(\,\mathrm{d}y) + \int_{\{x\}} \mu(\,\mathrm{d}y).$$

10 Let the distribution function F on $\mathbb{R}$ have a density f with respect to the Lebesgue measure. Prove that for the corresponding characteristic function ϕ one has $\phi(u) \to 0$ as $|u| \to \infty$. This result is known as the Riemann-Lebesgue lemma and its 'analytic counterpart' is of importance in harmonic analysis. You may assume additionally that f is continuos Lebesgue a.e. You will get a bonus point, if you prove the result for a general f (not necessarily continuous).

CHAPTER 3

Limit theorems

This chapter deals with a number of important limit theorems in probability theory. Their proofs are to a considerable extent based on characteristic function techniques.

3.1. Characteristic functions and weak convergence

In this section we study how characteristic functions relate to weak convergence.

Our first result says that weak convergence of probability measures implies pointwise convergence of their characteristic functions.

PROPOSITION 51. *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$ and let $\phi, \phi_1, \phi_2, \ldots$ be their characteristic functions. If $\mu_n \xrightarrow{w} \mu$, then $\phi_n(u) \to \phi(u)$ for every $u \in \mathbb{R}$.*

PROOF. Consider for fixed u the function $f(x) = e^{\mathrm{i}ux}$. It is obviously bounded and continuous and we obtain straight from the definition of weak convergence that $\mu_n(f) \to \mu(f)$. But $\mu_n(f) = \phi_n(u)$. □

PROPOSITION 52. *Let $\mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$. Let $\phi_1, \phi_2, \ldots$ be the corresponding characteristic functions. Assume that the sequence (μ_n) is tight and that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n\to\infty} \phi_n(u)$ exists. Then there exists a probability measure μ on $(\mathbb{R}, \mathcal{B})$, such that $\phi = \phi_\mu$ and $\mu_n \xrightarrow{w} \mu$.*

PROOF. Since (μ_n) is tight we use Prokhorov's theorem to deduce that there exists a weakly converging subsequence (μ_{n_k}) with a probability measure as limit. Call this limit μ. From Proposition 51 we know that $\phi_{n_k}(u) \to \phi_\mu(u)$ for all u. Hence we must have $\phi_\mu = \phi$. We will now show that any convergent subsequence of (μ_n) has μ as a limit. Suppose that there exists a subsequence $(\mu_{n'_k})$ with limit μ'. Then $\phi_{n'_k}(u)$ converges to $\phi_{\mu'}(u)$ for all u. But, since $(\mu_{n'_k})$ is a subsequence of the original sequence, by assumption the corresponding $\phi_{n'_k}(u)$ must converge to $\phi(u)$ for all u. Hence we conclude that $\phi_{\mu'} = \phi_\mu$ and then $\mu' = \mu$.

Suppose that the whole sequence (μ_n) does not converge to μ. Then there must exist a function $f \in C_b(\mathbb{R})$, such that $\mu_n(f)$ does not converge to $\mu(f)$. So, there is $\varepsilon > 0$, such that for some subsequence (n'_k) we have

$$|\mu_{n'_k}(f) - \mu(f)| > \varepsilon. \tag{3.1}$$

Using Prokhorov's theorem, the sequence $(\mu_{n'_k})$ has a further subsequence $(\mu_{n''_k})$ that has a limit probability measure μ''. By the same argument as above (convergence of the characteristic functions) we conclude that $\mu''(f) = \mu(f)$. Therefore $\mu_{n''_k}(f) \to \mu(f)$, which contradicts (3.1). □

Characteristic functions are a tool to give a rough estimate of the tail probabilities of a random variable, useful to establish tightness of a sequence of probability measures. To that end we will use the following lemma. By taking the complex conjugate, check first that $\int_{-a}^{a}(1-\phi(u))\,\mathrm{d}u \in \mathbb{R}$ for every $a > 0$.

LEMMA 53. *Let a random variable X have distribution μ and characteristic function ϕ. Then for every $K > 0$*

$$P(|X| > 2K) \leq K \int_{-1/K}^{1/K} (1-\phi(u))\,\mathrm{d}u. \tag{3.2}$$

PROOF. It follows from Fubini's theorem and

$$\int_{-a}^{a} e^{\mathrm{i}ux}\,\mathrm{d}u = 2\frac{\sin ax}{x}$$

that

$$\begin{aligned} K\int_{-1/K}^{1/K}(1-\phi(u))\,\mathrm{d}u &= K\int_{-1/K}^{1/K}\int (1-e^{\mathrm{i}ux})\,\mu(\mathrm{d}x)\,\mathrm{d}u \\ &= \int K\int_{-1/K}^{1/K}(1-e^{\mathrm{i}ux})\,\mathrm{d}u\,\mu(\mathrm{d}x) \\ &= 2\int\left[1-\frac{\sin(x/K)}{x/K}\right]\mu(\mathrm{d}x) \\ &\geq 2\int_{|x/K|>2}\left[1-\frac{\sin(x/K)}{x/K}\right]\mu(\mathrm{d}x) \\ &\geq \mu([-2K,2K]^c). \end{aligned}$$

since $\frac{\sin x}{x} \leq \frac{1}{2}$ for $x > 2$[1]. □

The following theorem is known as Lévy's continuity theorem.

THEOREM 54 (Lévy's continuity theorem). *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$ and $\phi_1, \phi_2, \ldots$ the corresponding characteristic functions. Assume that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n\to\infty}\phi_n(u)$ exists. If ϕ is continuous at zero, then there exists a probability measure μ on $(\mathbb{R}, \mathcal{B})$, such that $\phi = \phi_\mu$ and $\mu_n \xrightarrow{w} \mu$.*

PROOF. We will show that under the present assumptions, the sequence (μ_n) is tight. To this end we will use Lemma 53. Let $\varepsilon > 0$. Since ϕ is continuous at zero, the same holds for $\overline{\phi}$, and there is $\delta > 0$ such that $|\phi(u) + \phi(-u) - 2| < \varepsilon$ if $|u| < \delta$. Notice that $\phi(u) + \phi(-u)$ is real-valued and bounded from above by 2. Hence $2\int_{-\delta}^{\delta}(1-\phi(u))\,\mathrm{d}u = \int_{-\delta}^{\delta}(2-\phi(u)-\phi(-u))\,\mathrm{d}u \in [0, 2\delta\varepsilon)$.

By the convergence of the characteristic functions (which are bounded), the dominated convergence theorem implies that

$$\int_{-\delta}^{\delta}(1-\phi_n(u))\,\mathrm{d}u \to \int_{-\delta}^{\delta}(1-\phi(u))\,\mathrm{d}u.$$

Hence, for all $n \geq N$ with N chosen large enough, we have

$$\int_{-\delta}^{\delta}(1-\phi_n(u))\,\mathrm{d}u < 2\delta\varepsilon.$$

[1]Function $g(x) = (\sin x)/x$ is called the cardinal sine, or simply the sinc function.

It now follows from Lemma 53 that for $n \geq N$ and $K = 1/\delta$

$$\begin{aligned}\mu_n([-2K, 2K]^c) &\leq \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi_n(u))\,\mathrm{d}u \\ &< 2\varepsilon.\end{aligned}$$

We conclude that $(\mu_n)_{n\geq N}$ is tight and then so is the sequence $(\mu_n)_{n\in\mathbb{N}}$ as well. Apply Proposition 52 to conclude. □

COROLLARY 55. *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$ and ϕ, $\phi_1, \phi_2, \ldots$ be their corresponding characteristic functions. Then $\mu_n \xrightarrow{w} \mu$ if and only if $\phi_n(u) \to \phi(u)$ for all $u \in \mathbb{R}$.*

PROOF. If $\phi_n(u) \to \phi(u)$ for all $u \in \mathbb{R}$, then we can apply Theorem 54. Function ϕ, being a characteristic function, is continuous at zero. Hence there is a probability measure to which the μ_n weakly converge. But since the $\phi_n(u)$ converge to $\phi(u)$, the limiting probability measure must be μ. The converse statement we have encountered as Proposition 51. □

3.2. Weak law of large numbers

In this section we present the weak law of large numbers for a sequence of i.i.d. random variables. In its proof we will need the following elementary result from calculus.

LEMMA 56. *Let z be a complex number, such that $|z| \leq 1/2$. Then there exists a complex number θ depending on z, such that $|\theta| \leq 1$, and $\log(1+z) = z + \theta|z|^2$.*

PROOF. Without loss of generality, assume that $z \neq 0$ (when $z = 0$, $\log(1+z) = 0 = z$, and hence $\theta = 0$). We have

$$\begin{aligned}\log(1+z) &= z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4}\ldots \\ &= z + z^2\left(-\frac{1}{2} + \frac{z}{3} - \frac{z^2}{4} + \ldots\right) \\ &= z + |z|^2\frac{z^2}{|z|^2}\left(-\frac{1}{2} + \frac{z}{3} - \frac{z^2}{4} + \ldots\right).\end{aligned}$$

We claim that

$$\theta = \frac{z^2}{|z|^2}\left(-\frac{1}{2} + \frac{z}{3} - \frac{z^2}{4} + \ldots\right).$$

To verify the claim, we need to check that $|\theta| \leq 1$. This, however, is easy:

$$|\theta| = \left|-\frac{1}{2} + \frac{z}{3} - \frac{z^2}{4} + \ldots\right| \leq \frac{1}{2} + \frac{1}{3}\left(\frac{1}{2}\right) + \frac{1}{4}\left(\frac{1}{2}\right)^2 + \ldots \leq \sum_{k=1}^{\infty}\left(\frac{1}{2}\right)^k = 1.$$

□

COROLLARY 57. *If a sequence of complex numbers $\{c_n\}$ converges to the limit c, then*

$$\lim_{n\to\infty}\left(1 + \frac{c_n}{n}\right)^n = e^c.$$

PROOF. It is sufficient to prove that

$$\lim_{n\to\infty} \log\left(\left\{1+\frac{c_n}{n}\right\}^n\right) = \lim_{n\to\infty}\left\{n\log\left(1+\frac{c_n}{n}\right)\right\} = c.$$

Since the sequence $\{c_n\}$ converges, it is bounded, and furthermore, $|c_n/n| \leq 1/2$ for all n large enough. Then from Lemma 56,

$$\left\{n\log\left(1+\frac{c_n}{n}\right)\right\} = c_n + o(1).$$

Because the right-hand side tends to c as $n\to\infty$, the result follows. □

THEOREM 58 (Weak law of large numbers). *Let $X_1,\ldots,X_n$ be i.i.d. random variables with characteristic function ϕ. Assume that ϕ is differentiable at zero and $\phi'(0) = \imath\mu$. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mu.$$

PROOF. By differentiability of ϕ at zero, we have

$$\begin{aligned}\phi(t) &= \phi(0) + \phi'(0)t + o(t)\\ &= 1 + \imath\mu t + o(t).\end{aligned}$$

By independence of X_i's, for every fixed t,

$$\mathbb{E}\left[e^{\imath t\overline{X}_n}\right] = \phi^n\left(\frac{t}{n}\right) = \left(1+\imath\mu\frac{t}{n}+o\left(\frac{1}{n}\right)\right)^n.$$

As $n\to\infty$, by Corollary 57 the right-hand side converges to $e^{\imath t\mu}$. Now $\phi(t) = e^{\imath t\mu}$ is the characteristic function of a constant random variable μ. By Lévy's continuity theorem, $\overline{X}_n \rightsquigarrow \mu$. Since the convergence in distribution and in probability are equivalent for constant limits, it follows that $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$. □

REMARK 59. If $\mathbb{E}[|X_1|] < \infty$, then the dominated convergence theorem allows one to interchange the order of differentiation and expectation to obtain

$$\phi'(t) = \frac{\mathrm{d}}{\mathrm{dt}}\mathbb{E}\left[e^{itX_1}\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{dt}}e^{itX_1}\right] = \imath\mathbb{E}\left[X_1e^{itX_1}\right]. \tag{3.3}$$

For $t=0$ this yields $\phi'(0) = \imath\mathbb{E}[X_1] = \imath\mu$ and $\overline{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$, which is hardly surprising in light of the strong law of large numbers. However, integrability of X_1 is only a sufficient, but not a necessary condition to justify (3.3). Hence the weak law of large numbers holds under a weaker condition than the strong law. □

REMARK 60. The condition $\phi'(0) = \imath\mu$ is also necessary for convergence $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$. We will not prove this fact. For the proof see e.g. Theorem 2.5.5 in Révész (1968). An alternative necessary and sufficient condition for the weak law of large numbers, that does not employ characteristic functions, is also known (see e.g. Chung (2001), pp. 116–118). Furthermore, Chung (2001), pp. 118–119, contains an example, in which the weak law of large numbers holds, while the strong law fails.

3.3. Probabilities of large deviations

The weak law of large numbers does not provide information on the probabilities of large deviations of $\overline{X}_n$ from μ. Derivation of results in this setting is a task of an important and deep branch of probability theory, the large deviations theory. The latter is beyond the scope of the present course. We only remark that treatment of the case when a sequence of i.i.d. random variables $\{X_n\}$ satisfies Cramér's condition,

$$\exists \lambda > 0, \text{ s.t. } \varphi(\lambda) = \mathbb{E}[e^{\lambda X_1}] < \infty, \tag{3.4}$$

is relatively elementary and refer the reader to pp. 400–403 in Shiryaev (1996) for details. Under (3.4), $\mathbb{E}[X_1] = \mu < \infty$. The function φ is called the moment-generating function of X_1 or the Laplace transform (of the law) of X_1 (as it is often called in nonprobabilistic literature). It is obtained by replacing the argument of the characteristic function of X_1 with $-\imath\lambda$. In light of this the moment generating function possesses many properties similar to those of a characteristic function, but unlike the latter it does not always exist. Define the function ψ by $\psi(\lambda) = \log \varphi(\lambda)$ (this is the cumulant-generating function of X_1). The inequality one gets is

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \leq 2 \exp\left(-n \cdot \min(H(\mu - \varepsilon), H(\mu + \varepsilon))\right), \tag{3.5}$$

where the function

$$H(a) = \sup_{\lambda \in \mathbb{R}} [a\lambda - \psi(\lambda)]$$

is called the Cramér transform of X_1 (in terminology of convex analysis this is the Legendre transform of the cumulant-generating function ψ). The Cramér transform can be computed explicitly for a number of distributions, which yields explicit bounds on large deviations probabilities.

Example 61. Let $\{X_n\}$ be a sequence of i.i.d. Bernoulli random variables with probability of success $0 < p < 1$. Straightforward computations give that

$$H(a) = \begin{cases} a \log\left(\frac{a}{p}\right) + (1-a) \log\left(\frac{1-a}{1-p}\right) & \text{if } a \in [0,1], \\ \infty & \text{otherwise.} \end{cases}$$

Insert this expression in the right-hand side of (3.5) to obtain a bound on the probabilities of large deviations.

A much more crude bound on probabilities of large deviations is obtained by applying Chebyshev's inequality. If $\mathbb{V}[X_1] = \sigma^2$, then

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{\mathbb{V}[\overline{X}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

In particular, when $\{X_n\}$ is an i.i.d. sequence of Bernoulli random variables with probability of success p,

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}. \tag{3.6}$$

If we denote

$$p_n(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

the inequality (3.6) can be rewritten as

$$\sum_{\{k:|k/n-p|\geq\varepsilon\}} p_n(k) \leq \frac{1}{4n\varepsilon^2}.$$

We will use this fact to give a probabilistic proof of the Weierstraß theorem, which asserts that for any continuous function $u : [0,1] \to \mathbb{R}$ there exists a sequence of polynomials u_n, such that

$$\lim_{n\to\infty} \sup_{p\in[0,1]} |u_n(p) - u(p)| = 0, \tag{3.7}$$

see Theorem 7.26 in Rudin (1976). Take

$$u_n(p) = \sum_{k=0}^{n} u\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k}.$$

These are called Bernstein polynomials. We have

$$\mathbb{E}\left[u_n(\overline{X}_n)\right] = u_n(p).$$

Since the function u, being continuous on $[0,1]$, is uniformly continuous on that interval, for every $\varepsilon > 0$ one can find $\delta > 0$, such that $|u(x) - u(y)| \leq \varepsilon$, whenever $|x - y| \leq \delta$. Also note that u is bounded on $[0,1]$. We then get

$$\begin{aligned}
|u_n(p) - u(p)| &= \left|\sum_{k=0}^{n} \left[u\left(\frac{k}{n}\right) - u(p)\right] \binom{n}{k} p^k (1-p)^{n-k}\right| \\
&\leq \sum_{\{k:|k/n-p|\leq\delta\}} \left|u\left(\frac{k}{n}\right) - u(p)\right| p_n(k) \\
&+ \sum_{\{k:|k/n-p|\geq\delta\}} \left|u\left(\frac{k}{n}\right) - u(p)\right| p_n(k) \\
&\leq \varepsilon + \frac{\|u\|_\infty}{n\delta^2}.
\end{aligned}$$

The bound on the right-hand side is independent of p. Let $n \to \infty$ to obtain that the right-hand side of (3.7) does not exceed ε. Since ε is arbitrary, the result follows.

3.4. Central limit theorem

Let $\{X_i\}$ be a sequence of random variables. In general the distribution of the sum $S_n = \sum_{i=1}^n X_i$ might have a complicated form and hence be difficult to compute. The central limit theorem provides a simple approximation to it, that is very useful in practice.

Although the result holds in a much greater generality, we will prove the central limit theorem only for a sequence of i.i.d. random variables with finite second moments. The proof will yet again demonstrate the power of the method of characteristic functions.

THEOREM 62 (Central limit theorem). *Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $0 < \mathrm{Var}[X_i] = \sigma^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \rightsquigarrow N(0,1).$$

PROOF. Without loss of generality, we may suppose that $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}[X_i] = 1$ (otherwise replace X_i with $(X_i - \mu)/\sigma$, and note that this has mean 0 and variance 1). Let ϕ be the characteristic function of X_i. Since by assumption $\mathbb{E}[X_i^2] = 1$, the characteristic function ϕ is twice differentiable and (cf. p. 290 in Hardy (1967) and Proposition 37 (vi))

$$\begin{aligned}\phi(u) &= \phi(0) + \phi'(0)u + \phi''(0)\frac{u^2}{2} + o(u^2) \\ &= 1 - \frac{1}{2}u^2 + o(u^2).\end{aligned}$$

By independence of X_i's and Corollary 57 we then get for every fixed $t \in \mathbb{R}$ that

$$\begin{aligned}\mathbb{E}\left[e^{itS_n/\sqrt{n}}\right] &= \phi^n\left(\frac{t}{\sqrt{n}}\right) \\ &= \left\{1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{|t|}{\sqrt{n}}\right)^2\right\}^n \\ &= \left\{1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right\}^n \to e^{-t^2/2}.\end{aligned}$$

The limit being the characteristic function of a standard normal random variable Z, the proof is completed upon invoking Lévy's continuity theorem. □

EXAMPLE 63. Suppose we have an i.i.d. sample $X_1, \ldots, X_n$ from the Bernoulli distribution with probability of success p, but we do not know p. The parameter p can be estimated by the sample mean $\hat{p}_n = n^{-1}\sum_{i=1}^n X_i$. By the strong law of large numbers $\hat{p}_n \xrightarrow{a.s.} p$, and by the central limit theorem

$$\frac{\sqrt{n}}{\sqrt{p(1-p)}}(\hat{p}_n - p) \rightsquigarrow N(0,1).$$

Thus for large n the estimator $\hat{p}_n$ has approximately the normal distribution with mean p and variance $p(1-p)/n$, which gives an idea on the precision with which p is recovered as $n \to \infty$. The asymptotic variance $p(1-p)/n$ of the estimator can be estimated by $\hat{p}_n(1-\hat{p}_n)/n$, and by Slutsky's lemma

$$\frac{\sqrt{n}}{\sqrt{\hat{p}_n(1-\hat{p}_n)}}(\hat{p}_n - p) \rightsquigarrow N(0,1),$$

so that, roughly speaking, we do not need to know the value of p in order to determine the precision with which it is recovered by $\hat{p}_n$: by a somewhat circular argument the latter can be again estimated by using $\hat{p}_n$.

3.5. Delta method

Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $0 < \mathrm{Var}[X_i] = \sigma^2 < \infty$. By the central limit theorem,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \rightsquigarrow N(0,1).$$

Can we say something about the weak convergence of a sequence $\{g(\overline{X}_n)\}$, where $g : \mathbb{R} \mapsto \mathbb{R}$ is some fixed function? Such a question often arise in in statistics. When g is differentiable, the answer is given by the following result, known as the delta method.

THEOREM 64 (Delta method). *Assume that the conditions of the central limit theorem (Theorem 62) hold. Let g be differentiable at μ and $g'(\mu) \neq 0$. Then*

$$\sqrt{n}\frac{g(\overline{X}_n) - g(\mu)}{\sigma g'(\mu)} \rightsquigarrow N(0,1).$$

PROOF. The proof is an instance of an elegant application of the almost sure representation theorem. On some probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ there exist random variables

$$\widetilde{Z}_n \stackrel{d}{=} \frac{S_n - n\mu}{\sigma\sqrt{n}}, \quad \widetilde{Z} \sim N(0,1),$$

such that $\widetilde{Z}_n \xrightarrow{a.s.} \widetilde{Z}$ (under $\widetilde{\mathbb{P}}$). By the foregoing, the definition of a derivative, the facts that $\sigma\widetilde{Z}_n/\sqrt{n} \xrightarrow{a.s.} 0$ and $\widetilde{\mathbb{P}}(\widetilde{Z} \neq 0) = 1$, and the continuous mapping theorem we have

$$\begin{aligned}
\sqrt{n}\frac{g(\overline{X}_n) - g(\mu)}{\sigma g'(\mu)} &\stackrel{d}{=} \sqrt{n}\frac{g(\mu + \sigma\widetilde{Z}_n/\sqrt{n}) - g(\mu)}{\sigma g'(\mu)} \cdot 1_{[\widetilde{Z}_n \neq 0]} \\
&= \frac{g(\mu + \sigma\widetilde{Z}_n/\sqrt{n}) - g(\mu)}{\sigma\widetilde{Z}_n/\sqrt{n}} \cdot \frac{\sigma\widetilde{Z}_n}{\sigma g'(\mu)} \cdot 1_{[\widetilde{Z}_n \neq 0]} \\
&\xrightarrow{a.s.} g'(\mu) \cdot \frac{\sigma\widetilde{Z}}{\sigma g'(\mu)} \cdot 1_{[\widetilde{Z} \neq 0]}.
\end{aligned}$$

The last term is equal to $\widetilde{Z}$ ($\widetilde{\mathbb{P}}$-almost surely), whence it follows that

$$\sqrt{n}\frac{g(\overline{X}_n) - g(\mu)}{\sigma g'(\mu)} \rightsquigarrow N(0,1)$$

on the original probability space. □

EXAMPLE 65. This is a continuation of Example 63. Suppose we want to estimate the odds $r = p/(1-p)$. For example, if the data $X_1, \ldots, X_n$ are the outcomes of a medical treatment with $p = 3/4$, then a patient has odds $3:1$ of getting better. A natural estimator of r is $\hat{r}_n = \hat{p}_n/(1-\hat{p}_n)$, but how good is this estimator? Assume $0 < p < 1$. Firstly, by the strong law of large numbers and the continuous mapping theorem, $\hat{r}_n \xrightarrow{a.s.} r$. Secondly, by the delta method (take $g(p) = p/(1-p)$ in Theorem 64)

$$\frac{\sqrt{n(1-p)^3}}{\sqrt{p}}(\hat{r}_n - r) \rightsquigarrow N(0,1),$$

so that for large n the estimator $\hat{r}_n$ is approximately normally distributed with mean r and variance $p/[n(1-p)^3]$. The latter can be estimated by $\hat{p}_n/[n(1-\hat{p}_n)^3]$ and an application of Sutsky's lemma yields

$$\frac{\sqrt{n(1-\hat{p}_n)^3}}{\sqrt{\hat{p}_n}}(\hat{r}_n - r) \rightsquigarrow N(0,1).$$

3.6. Berry-Esseen theorem

Convergence of some quantity to a limit inevitably leads to the question of the rate of convergence. In the setting of the central limit theorem proved above, the question is this: let F_n be the distribution function of $(S_n - n\mu)/(\sigma\sqrt{n})$. Theorem 62 implies that for all $x \in \mathbb{R}$, $F_n(x) \to \Phi(x)$. Can we say something about the rate, at which the difference $|F_n(x) - \Phi(x)|$ converges to zero? A good estimate

on this quantity might be necessary in applications, in particular for numerical computations. One result in this direction is the Berry-Esseen theorem.

THEOREM 66 (Berry-Esseen theorem). *Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean zero, variance σ^2 and the third absolute moment $\gamma = \mathbb{E}[|X_i|^3] < \infty$. Then there exists a universal constant A_0, such that*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{A_0\gamma}{\sigma^3}\frac{1}{\sqrt{n}}.$$

We will not prove this theorem. For the proof see e.g. Chung (2001), Section 7.4 (that particular proof is based on the method of characteristic functions). The exact value of the constant A_0 is not known, but there exist good estimates on it (the latest (?) one is $A_0 \leq 0.5129$; this is quite sharp, because there also holds a lower bound proved by Esseen: $A_0 \geq (\sqrt{10}+3)/(6\sqrt{2\pi}) \approx 0.40973$). The estimate in Theorem 66 is 'generic'. For specific distributions, tighter bounds might hold. For instance, let $X_1, \ldots, X_n$ be jointly normal and i.i.d. with $X_i \sim N(0,1)$. Then $F_n = \Phi$ and $\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)|$ is in fact zero.

Exercises

1 Let $\{X_n\}$ be a sequence of random variables with $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{V}[X_n] < \infty$. Assume that the covariances $\mathrm{Cov}[X_i, X_j] \to 0$ as $|i-j| \to \infty$. Prove the following version of the law of large numbers (due to Bernstein):

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]\right| > \varepsilon\right) \to 0$$

as $n \to \infty$. Hint: a sequence of random variables $\{\xi_n\}$ converges to zero in probability, when both the mean $\mathbb{E}[\xi_n]$ and the variance $\mathbb{V}[\xi_n]$ converge to zero as $n \to \infty$ (show this).

2 Let $\{X_n\}$ be a sequence of i.i.d. random variables. Prove that $S_n = n^{-1/2}\sum_{i=1}^{n} X_i$ converges in probability as $n \to \infty$ if and only if $\mathbb{P}(X_1 = 0) = 1$.

3 Let $\{X_n\}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_1^2] < \infty$. Prove that

$$\frac{\max(|X_1|, \ldots, |X_n|)}{\sqrt{n}} \rightsquigarrow 0$$

as $n \to \infty$. Hint: for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{\max(|X_1|, \ldots, |X_n|)}{\sqrt{n}} \leq \varepsilon\right) = \left[\mathbb{P}(X_1^2 \leq n\varepsilon^2)\right]^n$$

and $n\varepsilon^2\mathbb{P}(X_1 > n\varepsilon) \to 0$ as $n \to \infty$.

4 Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean zero and variance one, and let $\{d_n\}$ be a sequence of nonnegative numbers, such that $d_n = o(D_n)$ for $D_n^2 = \sum_{i=1}^{n} d_i^2$. Prove that the sequence $\{d_n X_n\}$ satisfies the central limit theorem:

$$\frac{1}{D_n}\sum_{i=1}^{n} d_i X_i \rightsquigarrow Z$$

for $Z \sim N(0,1)$.

5 Let $\{X_n\}$ be a sequence of i.i.d. random variables, such that $\mathbb{P}(X_1 = \pm 1) = 1/2$. Set $S_i = \sum_{k=1}^{i} X_i$. Show that

$$\frac{1}{\sqrt{n}} \max_{1 \le i \le n} S_i \rightsquigarrow |Z|,$$

where $Z \sim N(0,1)$.

6 Let $\{X_n\}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$. Show that

$$\sqrt{n}\frac{\overline{X}_n}{\sigma_n} \rightsquigarrow N(0,1),$$

where

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Incidentally, the result of this exercise also shows that if Y_n possesses t-distribution with n degrees of freedom, then $Y_n \rightsquigarrow N(0,1)$. Explain why.

7 Let X_n have a $\mathrm{Bin}(n, \lambda/n)$ distribution (for $n > \lambda$). Show that $X_n \rightsquigarrow X$, where X has a Poisson(λ) distribution. This result is known as the Poisson theorem.

8 Let $X, X_1, X_2, \ldots$ be a sequence of random variables and Y a $N(0,1)$-distributed random variable independent of that sequence. Let ϕ_n be the characteristic function of X_n and ϕ that of X. Let p_n be the density of $X_n + \sigma Y$ and p the density of $X + \sigma Y$.

(i) If $\phi_n \to \phi$ pointwise, then $p_n \to p$ pointwise.

(ii) Let $f : \mathbb{R} \to \mathbb{R}$ be bounded by B. Show that $|\mathbb{E}f(X_n + \sigma Y) - \mathbb{E}f(X + \sigma Y)| \le 2B \int (p(z) - p_n(z))^+ \, dz$.

(iii) Show that $|\mathbb{E}f(X_n + \sigma Y) - \mathbb{E}f(X + \sigma Y)| \to 0$ (with f bounded) if $\phi_n \to \phi$ pointwise.

(iv) Prove without referring to Corollary 55 that $X_n \rightsquigarrow X$ iff $\phi_n \to \phi$ pointwise (hint: one implication is straightforward, for the other the result of Exercise 1.1 is useful).

Bibliography

J. Bak and D. J. Newman. *Complex analysis.* Third edition. Undergraduate Texts in Mathematics. Springer, New York, 2010.

P. Billingsley. *Weak Convergence of Measures: Applications in Probability.* Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 5. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1971.

K. L. Chung. *A Course in Probability Theory.* Third edition. Academic Press, Inc., San Diego, CA, 2001.

G. H. Hardy. *A Course of Pure Mathematics.* Tenth edition. Cambridge University Press, Cambridge, 1967.

K. R. Parthasarathy. *Probability measures on metric spaces.* Reprint of the 1967 original. AMS Chelsea Publishing, Providence, RI, 2005.

Yu. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.*, 1(2), 157–214, 1956.

S. I. Resnick. *A Probability Path.* Birkhäuser Boston, Inc., Boston, MA, 1999.

P. Révész. *The Laws of Large Numbers.* Academic Press, New York, 1968.

W. Rudin. *Principles of Mathematical Analysis.* Third edition. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976.

A. N. Shiryaev. *Probability.* Translated from the first (1980) Russian edition by R. P. Boas. Second edition. Graduate Texts in Mathematics, 95. Springer-Verlag, New York, 1996.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, 1998.