

Lecture 1: Weak Convergence Basics

October 10, 2019

Lecturer: Burak Büke

1 Motivation: Averaging Principles

By far the two most important theorems in probability theory are the Law(s) of Large Numbers (LLN) and the Central Limit Theorem (CLT). These results mainly state that the sample average exhibit a statistical regularity as the number of samples increases to infinity.

A basic form of the law of large numbers tells us that the sample average of independent and identically distributed (i.i.d.) observations approach to the true mean as the sample size tends to infinity. The LLN can be stated in two forms depending on the mode of convergence:

Theorem 1.1 (Weak Law of Large Numbers). *If $\{X_n\}_{n=1}^\infty$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_1] = \mu < \infty$, then for any $\epsilon > 0$*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| < \epsilon \right) = 1.$$

Theorem 1.2 (Strong Law of Large Numbers). *If $\{X_n\}_{n=1}^\infty$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_1] = \mu < \infty$,*

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} \left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| = 0 \right) = 1.$$

Theorem 1.1 is a conclusion about convergence of probability, but it does not say anything about convergence of actual sequences. The convergence of actual realizations is dealt by the stronger version in Theorem 1.2. The LLN concludes the convergence, but it does not say anything about the *rate of convergence*. The CLT somewhat deals with the rate of convergence by characterizing how much the sample average deviates from the expected value.

Theorem 1.3 (Central Limit Theorem). *If $\{X_n\}_{n=1}^\infty$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_1] = \mu < \infty$ and $\mathbb{E}[(X_1 - \mu)^2] = \sigma^2 < \infty$, then the sample average $(X_1 + \cdots + X_n)/n$ converges “in distribution” to a normal random variable with mean μ and variance σ^2/n .*

Intuitively speaking, CLT states that the difference between sample mean and the expected value is distributed normally at a scale of $1/\sqrt{n}$. The conditions stated in the above theorems are a bit restrictive and they can be significantly relaxed. In both LLN and CLT the result is independent of the actual distribution of the underlying sequence and for this reason this type of results are sometimes referred as *invariance principles*. The concept of “convergence in distribution” concluded in Theorem 1.3 is currently left a bit vague and will be the main topic of these lectures. We will cover how the above results can be generalized and applied to more general situations, especially to stochastic processes. Our main tools will be those of real and functional analysis. We now start with reviewing basic probability concepts to introduce our terminology and notation.

2 Review of Basic Probability Concepts

2.1 The Probability Triple

To study probability in a formal manner, we need to first define an appropriate space on which we define our random quantities. We take Ω to be our sample space which denote the outcome of a random experiment. A collection of subsets of Ω , $\mathcal{F} \subset 2^\Omega$, is a σ -algebra if it includes Ω and it is closed under complements, countable unions (and hence countable intersections). We define the probability measure $\mathbb{P}$ as a map from $\mathcal{F}$ to the interval $[0, 1]$ such that

1. For any $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$ and $\mathbb{P}(\Omega) = 1$.
2. For any countable disjoint sequence $A_1, A_2, \dots \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is generally called a probability triple or a probability space.

Example 2.1. For the random experiment of flipping a fair coin, we take $\Omega = \{T, H\}$, $\mathcal{F} = \{\emptyset, \{T\}, \{H\}, \{T, H\}\}$ and $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{T\}) = 0.5$, $\mathbb{P}(\{H\}) = 0.5$, $\mathbb{P}(\{T, H\}) = 1$.

Example 2.2. We can take $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{L}((0, 1))$, the family of Lebesgue measurable subsets of $[0, 1]$ and $\mathbb{P} = \mathbb{L}$, the Lebesgue measure.

We require the probability space to be large enough to support the random quantities that we investigate, but apart from that we will not be concerned with the exact structure of the probability space in general.

2.2 Random Variables, Random Elements and Expectations

The Borel σ -algebra on $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$, is defined to be the smallest σ -algebra that contains all open real intervals. A random variable is a measurable mapping from the sample space $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, i.e.,

Definition 2.3. $X : \Omega \rightarrow \mathbb{R}$ is a random variable if $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{B}(\mathbb{R})$.

A given random variable X defines a probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for any $A \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}),$$

which we refer as the probability distribution of X . The cumulative distribution of X is $F_X(x) = \mathbb{P}_X((-\infty, x])$.

Example 2.4. Using the probability triple $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1), \mathcal{L}((0, 1)), \mathbb{L})$, $X(\omega) = \omega$ defines a uniform(0,1) random variable.

Example 2.5. Using the probability triple $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1), \mathcal{L}((0, 1)), \mathbb{L})$, $X(\omega) = -\frac{\ln(\omega)}{\lambda}$ defines an exponential random variable with rate λ .

To verify Example 2.5 we can use the following proposition:

Proposition 2.6. *Let X be a random variable from a continuous distribution with strictly increasing cumulative $F(x)$. Let U be a uniform(0,1) random variable then $F^{-1}(U)$ follows $F(x)$.*

Proof. For any $y \in [0, 1]$, $F(y) = P(U \leq y) = y$. Then,

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x) = P(X \leq x).$$

□

Proposition 2.6 can be generalized for a random variable with any cumulative distribution function by defining the generalized inverse as

$$F^{-1}(u) = \inf\{y | F(y) \geq u\}.$$

If $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in Definition 2.3 is replaced with an abstract space $(E, \mathcal{E})$ then X is called a random element. Some examples of space E that we will see in this course will include the space of d -dimensional real vectors $\mathbb{R}^d$, the space of continuous functions, the space of right-continuous left-limit (càdlàg) functions.

The expectation of a random variable X is defined as

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

Similarly, the expectation of a function of X is

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

3 Modes of Convergence

As we have seen while investigating the laws of large numbers, it is possible to define convergence of random variables in different ways. We will be investigating four different types of convergence:

1. **Convergence in Probability (a.k.a. Convergence in Measure)** A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ is said to converge in probability to X , denoted $X_n \xrightarrow{p} X$, if for any $\epsilon > 0$

$$\mathbb{P}(|X_n - X| > \epsilon) \rightarrow 0.$$

2. **Almost Sure Convergence (a.k.a. Strong Convergence)** $\{X_n\}_{n=1}^{\infty}$ is said to converge almost surely, if the set

$$A = \{\omega \in \Omega : \lim_{n \rightarrow \infty} |X_n(\omega) - X(\omega)| \neq 0\}$$

is a zero probability event. We will also refer to this as “with probability 1 (w.p.1) convergence”.

3. **Convergence in L^p** $\{X_n\}_{n=1}^{\infty}$ is said to converge in L^p if

$$\mathbb{E}[|X_n - X|^p] \rightarrow 0.$$

4. **Convergence in Distribution (a.k.a. Weak Convergence)** This mode of convergence is the main topic of the course and we will provide a precise definition in the next section. Roughly speaking X_n converging to X imply that as $n \rightarrow \infty$ the distribution of X_n becomes more and more similar to the distribution of X .

The modes of convergence for random variables can be generalized to random elements that are taking values in a metric space E by replacing $|X_n - X|$ with the metric $d(X_n, X)$ on E .

3.1 Properties of Different Modes of Convergence

In this section, we present the relationship between different modes of convergence.

Proposition 3.1. *If $\{X_n\}_{n=1}^\infty$ converge almost surely, then they also converge in probability.*

Proof. $X_n \xrightarrow{a.s.} X$ if and only if for all $\epsilon > 0$ $\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$ w. p. 1. Hence, using Fatou's lemma

$$\begin{aligned} 0 = \int \lim_{n \rightarrow \infty} I(|X_n - X| > \epsilon) d\mathbb{P} &\geq \limsup_{n \rightarrow \infty} \int I(|X_n - X| > \epsilon) d\mathbb{P} \\ &= \limsup_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) \\ &= \liminf_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) \geq 0. \end{aligned}$$

□

Proposition 3.1 indicates that almost sure convergence is stronger than convergence in probability. Generally, convergence in probability does not imply almost sure convergence as can be seen in the following standard example.

Example 3.2. *Take $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1), \mathcal{L}((0, 1)), \mathbb{L})$ and if $k(k-1)/2 < n \leq (k+1)(k+1)/2$ define*

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \left(\frac{n-1-k(k-1)/2}{k}, \frac{n-k(k-1)/2}{k} \right) \\ 0 & \text{otherwise.} \end{cases}$$

For all $\epsilon > 0$ we have $\lim_{n \rightarrow \infty} \mathbb{P}(|X_n| > \epsilon) = 0$. However, the sequence $\{X_n(\omega), n \geq 0\}$ does not have a limit for any $\omega \in \Omega$ as it alternates between 0 and 1.

Even though convergence in probability does not imply almost sure convergence, probabilistic convergence still have some almost sure implications.

Proposition 3.3. *For any sequence $X_n \xrightarrow{p} X$ we can find a subsequence $X_{n(k)}$ such that $X_{n(k)}$ converges to X almost surely.*

Proof. Choose $n(k)$ such that $\mathbb{P}(|X_{n(k)} - X| > 2^{-k}) < 2^{-k}$ and use Borel-Cantelli Lemma (see e.g. [3, Theorem 2.18]). □

4 Convergence in Distribution in $\mathbb{R}$

Now, we are to define what is meant by a sequence of random variables $\{X_n\}_{n=1}^\infty$ converging in distribution to X , which we generally denote as $X_n \Rightarrow X$. The first obvious candidate is to require

$$\mathbb{P}(X_n \in A) \rightarrow \mathbb{P}(X \in A) \text{ for all } A \in \mathcal{F},$$

which is equivalent to requiring $F_{X_n}(x) \rightarrow F_X(x)$. However, this is too strong for accommodating most cases. For example, consider the deterministic sequence $X_n = a_n$ w.p.1, where the real sequence $\{a_n\}_{n=1}^\infty$ converges to a and $a_n > a$ for all n . We generally would want to say that X_n converges in distribution to the deterministic $X = a$, but $F_{X_n}(a) = 0 \not\rightarrow F_X(a) = 1$. This motivates the following definition for the convergence in distribution:

Definition 4.1. A sequence of random variables $\{X_n\}_{n=1}^\infty$ converges in distribution to random variable X if and only if $F_{X_n}(x) \rightarrow F_X(x)$ for all continuity points of $F_X(\cdot)$.

There is an equivalent definition of convergence in distribution using bounded continuous functions and expectations.

Theorem 4.2. $X_n \Rightarrow X$ if and only if $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$ for all $g \in C_b(\mathbb{R})$, the set of bounded continuous functions.

A general version of this theorem is proven below as a part of Portmanteau (Alexandrov's) Theorem and hence we omit it here.

If we have a infinite collection of a collection of cumulative distributions, then we can always find a subsequence that converges to a function using a diagonalization argument.

Theorem 4.3 (Helly's Selection Theorem). For any sequence of cumulative distributions $\{F_{X_n}\}_{n=1}^\infty$, there exists a subsequence $\{F_{X_{n(k)}}\}_{k=1}^\infty$ that converges to a right-continuous function F at its all continuity points.

Proof. To use the diagonalization argument, first consider an enumeration of the rational numbers $r_1, r_2, \dots$. Then, fixing r_1 we have $0 \leq F_{X_n}(r_1) \leq 1$ for all n which implies that it has a subsequence $n_1^1, n_2^1, \dots$ where $\{F_{X_{n_i^1}}(r_1)\}_{i=1}^\infty$ converges to $F(r_1)$. Then, only considering this subsequence, we fix r_2 and find a further subsequence $n_1^2, n_2^2, \dots$ where $\{F_{X_{n_i^2}}(r_2)\}_{i=1}^\infty$ converges to $F(r_2)$. Iterating this procedure, we can find a subsequence $n_1^\infty, n_2^\infty, \dots$ which converges for all rational numbers to the right-continuous increasing function

$$F(x) = \inf_{x \leq r \in Q} F(r).$$

To conclude the proof, we need to prove that $\{F_{X_{n_i^\infty}}(x)\}$ converges to $F(x)$ at all continuity points. Suppose x is a continuity point of F such that $\tilde{r}_1 < x < \tilde{r}_2$ and $|F(\tilde{r}_i) - F(x)| < \epsilon/4$ for $i = 1, 2$, then for large enough i

$$|F_{X_{n_i^\infty}}(x) - F(x)| \leq |F_{X_{n_i^\infty}}(x) - F_{X_{n_i^\infty}}(\tilde{r}_2)| + |F(\tilde{r}_2) - F(x)| + |F(\tilde{r}_2 - F_{X_{n_i^\infty}}(\tilde{r}_2))| < \epsilon.$$

□

Helly's selection theorem concludes convergence of a subsequence to a non-decreasing and right-continuous function F , but this limiting function need not be a distribution function as $F(x)$ may not approach to 1 as $x \rightarrow \infty$. To see this again consider a series of random variables $X_n = a_n$ w.p.1 and $a_n \rightarrow \infty$ as $n \rightarrow \infty$. This implies $F_{X_n}(x) \rightarrow 0$ for all x , which is right-continuous and non-decreasing, but is not a cumulative distribution. In other words, for distribution F_{X_n} , all the probability mass is concentrated on the point a_n and this mass escapes to infinity as $n \rightarrow \infty$. The concept of tightness helps us to characterize when the limiting function is a cumulative distribution.

Definition 4.4. A family of random variables $\{X_\alpha\}_{\alpha \in S}$ is tight if and only if for every $\epsilon > 0$ there exists $M_\epsilon > 0$ such that

$$\sup_{\alpha \in S} \mathbb{P}(X_\alpha \in \mathbb{R} \setminus [-M_\epsilon, M_\epsilon]) < \epsilon.$$

Theorem 4.5. Suppose that the sequence of cumulative distributions F_{X_n} converges to F at every continuity point of F . Then, F is a cumulative distribution if and only if the sequence of corresponding random variables $\{X_n\}_{n=1}^\infty$ is tight.

This is a special case of Prokhorov's theorem that we discuss in detail below and hence, we omit the proof here.

4.1 Characteristic Functions

Characteristic functions (not to be confused with the same term used in analysis) provide us with useful tools of Fourier analysis for characterizing the distributions and their convergence. Characteristic function $\varphi_X(t)$ of a random variable X is defined as

$$\varphi_X(t) := \mathbb{E}[e^{itX}]$$

and there is a one-to-one correspondence between convergence of characteristic functions and convergence in distribution. The proof of the following theorem follows the steps from Durrett [2].

Theorem 4.6 (Continuity Theorem). Let X_n be a sequence of random variables with characteristic functions φ_{X_n} , then X_n converges in distribution to a random variable X with characteristic distribution φ_X if and only if $\varphi_{X_n}(t) \rightarrow \varphi_X(t)$ for all $t \in \mathbb{R}$ and $\varphi_X(t)$ is continuous at 0.

To prove this theorem, we need the following two lemmas. The first one is a basic result in topological spaces and the second relates probability to characteristic functions.

Lemma 4.7. Let a_n be a sequence in a topological space. If every sequence has a subsequence that converges to a , then $a_n \rightarrow a$.

Proof. Suppose that $a_n \not\rightarrow a$, then this implies that there is a neighbourhood of a which excludes an infinite number of elements. This implies that no subsequence out of these infinite elements converges to a , which is a contradiction. $\square$

Lemma 4.8. Consider the random variable X with probability distribution $\mathbb{P}_X$ with characteristic function φ . We have the following relationship

$$\mathbb{P}_X(|X| > M) \leq \frac{M}{2} \int_{-2/M}^{2/M} (1 - \varphi(t)) dt$$

Proof. First, consider the righthand side of the equation

$$\begin{aligned}
\frac{M}{2} \int_{-2/M}^{2/M} (1 - \int e^{itx} \mathbb{P}(dx)) dt &= \frac{M}{2} \left(\frac{4}{M} - \int_{-2/M}^{2/M} \int e^{itx} \mathbb{P}(dx) dt \right) \\
&= 2 - \frac{M}{2} \int \int_{-2/M}^{2/M} e^{itx} dt \mathbb{P}(dx) \\
&= 2 - \frac{M}{2} \int \left(\int_{-2/M}^{2/M} (\cos(tx) + i \sin(tx)) dt \right) \mathbb{P}(dx) \\
&= 2 - 2 \int \frac{\sin(2x/M)}{2x/M} \mathbb{P}(dx) \\
&= 2 \int \left(1 - \frac{\sin(2x/M)}{2x/M} \right) \mathbb{P}(dx)
\end{aligned}$$

Now, using the fact that $|\sin(x)| \leq |x|$, we can discard the interval $(-M, M)$ and get

$$\begin{aligned}
\frac{M}{2} \int_{-2/M}^{2/M} (1 - \int e^{itx} \mathbb{P}(dx)) dt &\geq 2 \int_{|x| \geq M} \left(1 - \frac{1}{2x/M} \right) \mathbb{P}(dx) \\
&\geq \mathbb{P}(|X| > M).
\end{aligned}$$

□

Proof of Theorem 4.6. Using the dominated convergence theorem $\mathbb{P}_{X_n}(|X_n| > M) \rightarrow 0$. Hence, Theorem 4.5, the probability distributions are tight and each sequence has a subsequence converging to the distribution uniquely determined by φ . Then, using Lemma 4.1 implies the stated result. □

The proof of Theorem 4.6 follows the standard steps in proving the weak convergence outlined as follows:

1. Determine a way to uniquely characterize the probability measures of interest. In this case, it is the characteristic functions.
2. Prove that the collection of probability measures under consideration is tight, this implies that each sequence has a converging subsequence.
3. Use your characterization to prove that any subsequence that converges converges to the same limit.

In the basic example of $\mathbb{R}$ above, characteristic functions are enough to perform all three steps. However, in the general case, we may need to use different methods at different steps.

As a final remark, we state the basic version of the central limit theorem in Theorem 1.3. Note that the assumptions are rather restrictive to allow for a short proof and can be relaxed significantly.

Proof of Theorem 1.3: Using the transformation $(X_n - \mu)/\sigma$, without loss of generality we can assume that for our random variables the expected value is 0 and variance is 1. The relation $|e^{it} - 1| \leq t$ allows us to interchange the derivative and integral in the definition of characteristic functions and write

$$\frac{d^k \varphi(t)}{dt^k} = \mathbb{E}[(iX_n)^k e^{itX_n}].$$

Now, writing the Taylor expansion at $t = 0$, we get

$$\varphi_{X_n}(t) = 1 + i\mathbb{E}[X_n] - \frac{\mathbb{E}[X_n^2]t^2}{2} + o(t^2) = 1 - \frac{t^2}{2} + o(t^2).$$

Note that $\varphi_{(X_1+\dots+X_n)/\sqrt{n}} = \varphi(t/\sqrt{n})^n$. Hence

$$\varphi_{(X_1+\dots+X_n)/\sqrt{n}} = (1 - \frac{t^2}{2} + o(t^2))^n \rightarrow e^{-t^2/2}.$$

Using the fact that the characteristic function of a standard normal random variable Z is $\varphi_Z(t) = e^{-t^2/2}$, the result follows from Theorem 4.6. $\square$

Exercise 4.9. Prove that $\varphi_Z(t) = e^{-t^2/2}$.

5 Probability Measures on Metric Spaces

In this section, we begin generalizing the concept for metric spaces. This part closely follows Chapter 1 of Billingsley [1]. In this section, we consider the probability measures defined on metric space S equipped with metric ρ and the Borel σ -algebra $\mathcal{B}(S)$. Considering all sets in the σ -algebra will generally be tedious and it will be a lot easier to restrict our attention only to open and closed sets. To do so, we need the “regularity” of the measures.

Theorem 5.1. *On a metric space $(S, \mathcal{B}(S))$, any bounded measure μ is regular, i.e., for any $A \in \mathcal{B}(S)$ and $\epsilon > 0$, there exists a closed set F and open set G such that $F \subset A \subset G$ and $\mu(G \setminus F) < \epsilon$.*

Proof. First, we define the set of functions such that $\mathcal{A} = \{A \in \mathcal{B}(S) : A \text{ is regular}\}$. We first show that $\mathcal{A}$ is a σ -algebra. Notice that $S \in \mathcal{A}$ and $A \in \mathcal{A}$ implies that there exists F, G as suggested. Then, $G^c \subset A^c \subset F^c$ and $F^c \setminus G^c = G \setminus F$, hence $A^c \in \mathcal{A}$. We need to prove that union of a countable collection of sets $\{A_i\}_{i=1}^\infty \subset \mathcal{A}$ is also in $\mathcal{A}$, i.e., $A = \bigcup_{i=1}^\infty A_i \in \mathcal{A}$. First, we can see that $B_n := \bigcup_{i=1}^n A_i \in \mathcal{A}$. Fix $\epsilon > 0$. We choose closed $F_i \subset A_i$ and open set $G_i \supset A_i$ such that $\mu(A_i \setminus F_i) = \epsilon/2N$ and $\mu(G_i \setminus A_i) < \epsilon/2N$ and set $\tilde{F}_n = \bigcup_{i=1}^n F_i$ and $\tilde{G}_n = \bigcup_{i=1}^n G_i$, then $\tilde{F}_n \subset B_n \subset \tilde{G}_n$ and $\mu(\tilde{G}_n \setminus \tilde{F}_n) < \epsilon$. For countable collections, we use continuity from below for bounded measures (c.f. Exercise 5.2) and set $N > 0$ such that $\mu(A \setminus B_N) < \epsilon/2$. Now, we can find two closed sets F^1, F^2 such that $\mu(B_N \setminus F^1) < \epsilon/2$ and $\mu(B_N^c \setminus F^2) < \epsilon/2$. Then, $F = F^1$ and $G = (F^2)^c$ satisfy the condition for regularity of A .

Now, we prove that any closed set F is in $\mathcal{A}$ and hence $\mathcal{A} = \mathcal{B}(S)$. We define open sets

$$G^\delta := \{y \in S : \exists x \in F \text{ with } \rho(x, y) < \delta\}. \quad (1)$$

Using the continuity of bounded measures from below $\mu(G^\delta) \rightarrow \mu(F)$ as $\delta \rightarrow 0$. Hence, we can find an F and G such that $\mu(G \setminus F) < \epsilon$ and this proves the theorem. $\square$

Exercise 5.2. (i) (Continuity from Below.) Suppose μ is a bounded measure and $B_n \uparrow B$, i.e., $B_1 \subset B_2 \subset \dots$ and $B_n \rightarrow B$. Show that $\mu(B_n) \rightarrow \mu(B)$.

(ii) (Continuity from Above.) Suppose μ is a bounded measure and $B_n \downarrow B$, i.e., $B_1 \supset B_2 \supset \dots$ and $B_n \rightarrow B$. Show that $\mu(B_n) \rightarrow \mu(B)$.

Theorem 5.1 states that on a metric space, it is possible to characterize a bounded measure only by looking at its values closed (or open) sets. This suggests that the following definition:

Definition 5.3. A collection $\mathcal{A}$ of subsets of S is called a separating class if $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{A}$ implies $\mu_1 = \mu_2$, i.e., $\mu_1(B) = \mu_2(B)$ for all $B \in \mathcal{B}(S)$.

Using this terminology the class of all closed subsets and the class of all open subsets of S are separating classes. We now use this fact to characterize a probability measure in terms of bounded continuous functions.

Theorem 5.4. Two bounded measures μ_1 and μ_2 coincide on $\mathcal{B}(S)$ if $\int f(x)\mu_1(dx) = \int f(x)\mu_2(dx)$ for all bounded, uniformly continuous functions.

Proof. For any closed set F , we define

$$f_F(x) = (1 - \inf\{\rho(x, y) : y \in F\}/\delta)^+. \quad (2)$$

This function takes value 1 for $x \in F$, 0 for $x \in (F^\delta)^c$ and is between 0 and 1 for $x \in F^\delta \setminus F$. For any closed F $\mu_1(F) \leq \int f_F(x)\mu_1(dx) = \int f_F(x)\mu_2(dx) = \mu_2(F^\delta)$. Taking $\delta \rightarrow 0$, $\mu_1(F) \leq \mu_2(F)$ for all closed sets F . The theorem follows using symmetry. $\square$

In Section 4, we introduced the concept of tightness of a collection of real random variables. We can generalize this concept for general metric spaces by replacing the bounded interval $[-M_\epsilon, M_\epsilon]$ with a compact set.

Definition 5.5. A family of probability measures $\{P_\alpha\}_{\alpha \in I}$ on metric space $(S, \mathcal{B}(S))$ is tight if and only if for any $\epsilon > 0$ there exists a compact set K_ϵ such that

$$\sup_{\alpha \in I} \mathbb{P}_\alpha(K_\epsilon^c) < \epsilon.$$

An interesting question is whether a single probability measure defined on an arbitrary metric space is tight and the following theorem deals with this.

Theorem 5.6. If the metric space $(S, \mathcal{B}(S))$ is complete and separable, then any probability measure on this space is tight.

Proof. For each k , we can cover the space S with countably many open balls of radius $1/k$, which we denote as $B_1^k, B_2^k, \dots$. Again using continuity from below, we can find an n_k such that

$$\mathbb{P}\left(\bigcup_{i=1}^{n_k} A_i^n\right) \geq 1 - \epsilon/2^{-k},$$

and

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \left(\bigcup_{i=1}^{n_k} A_i^n\right)^c\right) \leq \sum_{k=1}^{\infty} \mathbb{P}\left(\left(\bigcup_{i=1}^{n_k} A_i^n\right)^c\right) = \epsilon.$$

Using the fact that $\bigcap_{k=1}^{\infty} \bigcup_{i=1}^{n_k} A_i^n$ is totally bounded (can be covered by finitely many balls of radius ϵ for any ϵ), its closure say is compact (c.f. Theorem A.4 in Rudin [4]) and has probability greater than $1 - \epsilon$. $\square$

Example 5.7. Let $S = \mathbb{R}^\infty$, the set of all sequences $x = \{x_i\}_{i=1}^\infty$ and endow it with the metric

$$\rho(x, y) = \sum_{i=1}^{\infty} \frac{1 \wedge |x_i - y_i|}{2^i}.$$

Using this metric $x^n \rightarrow x$ if and only if $x_i^n \rightarrow x_i$ for all i . With this metric, $\mathbb{R}^\infty$ is also separable and complete (why?).

Dealing with infinite dimensions is rather difficult and hence if we can get away with analyzing finite dimensions, we wish to proceed doing so. First, we recall the properties of the base of a topology.

Definition 5.8. A collection of open sets $\mathcal{T}$ is a basis for a topology if and only if each open set G contains an element of $\mathcal{T}$.

Theorem 5.9. A collection of open subsets of S , $\mathcal{T}$ is called the basis for a topology if every open subset of S can be written as the union of members of $\mathcal{T}$. If the underlying space is separable, then this can be done with countably many sets in $\mathcal{T}$.

Now, we define $N_{k,\epsilon}(x) = \{y : |x_i - y_i| < \epsilon \text{ for all } 1 \leq i \leq k\}$. The collection of sets of the form $N_{k,\epsilon}(x)$ is a basis for the topology on $\mathbb{R}^\infty$ imposed by the suggested metric. To see this, for any open ball around x with radius ϵ , we can select k such that $\sum_{i=k}^{\infty} 1/2^i < \epsilon/2$ and $N_{k,\epsilon/2k}(x)$ to lie in the open ball.

Now, the σ -algebra generated by sets of the form $N_{k,\epsilon}(x)$ should be equal to $\mathcal{B}(S)$ and hence, in general it is enough to consider only finite dimensional sets to characterize any probability measure in $\mathbb{R}^\infty$.

5.1 Aleksandrov's Theorem (Portmanteau)

For real numbers, we defined convergence in distribution as convergence of the cumulative distributions at all continuity points of the limiting distribution and related it to the convergence of expected values of bounded continuous functions. Now, we define convergence in measure (distribution) in more general metric spaces.

Definition 5.10. A sequence of probability measures on $(S, \mathcal{B}(S))$, $\{\mathbb{P}_n\}_{n=1}^\infty$ is said to converge in measure to $\mathbb{P}$, denoted $\mathbb{P}_n \Rightarrow \mathbb{P}$, if for all bounded continuous real-valued functions f

$$\int_S f(x) \mathbb{P}_n(dx) \rightarrow \int_S f(x) \mathbb{P}(dx).$$

As we promised above, it is possible to relate convergence in measure to convergence of probabilities of events. Aleksandrov's Theorem, which is also to as Portmanteau Theorem as it combines various aspects, establishes this relationship.

Theorem 5.11 (Aleksandrov's Theorem). For a sequence of probability measures $\{\mathbb{P}_n\}_{n=1}^\infty$ the following are equivalent:

- (i) $\mathbb{P}_n \Rightarrow \mathbb{P}$
- (ii) $\int_S f(x) \mathbb{P}_n(dx) \rightarrow \int_S f(x) \mathbb{P}(dx)$ for all "uniformly" continuous bounded f .

(iii) For all closed F , $\limsup_{n \rightarrow \infty} \mathbb{P}_n(F) \leq \mathbb{P}(F)$.

(iv) For all open G , $\liminf_{n \rightarrow \infty} \mathbb{P}_n(F) \geq \mathbb{P}(F)$.

(v) For all continuity sets A , i.e., boundary of the set has 0 probability $\mathbb{P}(\partial A) = 0$, $\lim_{n \rightarrow \infty} \mathbb{P}_n(A) = \mathbb{P}(A)$.

Proof. By definition, (i) implies (ii). To prove (ii) $\rightarrow$ (iii) for any closed set F consider function $f_F(x)$ and open set G^δ defined in equations (2) and (1), respectively. Then,

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n(F) \leq \lim_{n \rightarrow \infty} \int_S f(x) \mathbb{P}_n(dx) = \int_S f(x) \mathbb{P}(dx) \leq \mathbb{P}(G^\delta).$$

Again, using continuity from above, the claim follows by $\delta \rightarrow 0$. (iii) $\rightarrow$ (iv) is a direct consequence of open sets being complements of closed sets. Now we show that (iii) and (iv) together imply (v). For a $\mathbb{P}$ continuity set A , consider its closure $\bar{A}$ and interior A° . From the definition of $\mathbb{P}$ -continuity $\mathbb{P}(\bar{A}) = \mathbb{P}(A^\circ) = \mathbb{P}(A)$. Then,

$$\mathbb{P}(A) = \mathbb{P}(\bar{A}) \geq \limsup_{n \rightarrow \infty} \mathbb{P}_n(\bar{A}) \geq \liminf_{n \rightarrow \infty} \mathbb{P}(A^\circ) \geq \mathbb{P}(A^\circ) = \mathbb{P}(A).$$

Now, we prove the final relation (v) $\rightarrow$ (i). Writing any bounded continuous function as $f = f^+ - f^-$, we can concentrate on functions $f \geq 0$. Now, remember that the expectation of any positive random variable X can be written as

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) dt.$$

Hence,

$$\int_S f(x) \mathbb{P}(x) = \int_0^C \mathbb{P}(f > t) dt,$$

where C is the upper bound of f . The boundary of sets $\partial\{x \in S : f(x) > t\} \subset \{x \in S : f(x) = t\}$ which is countable and using dominated convergence theorem the result follows. $\square$

6 Prokhorov's Theorem

While studying convergence in distribution (weak convergence) for probability measures in $\mathbb{R}^n$, we have seen that the ability to characterize whether each sequence has a convergent subsequence is of central importance and associated this with analyzing if we have any probability mass disappearing as $n \rightarrow \infty$. Famous Prokhorov's theorem formally associates these two concepts for general metric spaces.

Definition 6.1. A collection of probability measures $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ is relatively compact if and only if every infinite collection has a convergent subsequence.

Definition 6.2. A collection of probability measures $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ is tight if and only if for any $\epsilon > 0$, one can find a compact $K \subset S$ such that $\sup_{\alpha \in \mathcal{A}} \mathbb{P}(K^c) < \epsilon$.

Theorem 6.3 (Prokhorov's Theorem). On a metric space S ,

(i) if a collection of probability measures $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ is tight, then it is relatively compact.

(ii) Suppose S is also separable and complete. If $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ is relatively compact then it is tight.

Our proof of Prokhorov's Theorem follows Billingsley [1] rather closely.

Proof. The tightness of $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ imply that there exists a sequence of compact sets such that $K_1 \subset K_2 \subset \dots$, and $\sup_{\alpha \in \mathcal{A}} \mathbb{P}_\alpha(K_m^c) < 1/m$. Fix m , and consider the class of open balls

$$\mathcal{C}_i = \{B(x, 1/i) : x \in K_m\}.$$

For each i class covers K_m and compactness implies that there is a finite subcover of N_i^m balls centered at points $\{x_k^{i,m}\}_{k=1}^{N_i^m}$. The countable set $\{x_k^{i,m} : k = 1 \dots, N_i, i \in \mathbb{Z}_+\}$ is dense in K_m and this set is separable. This implies that $\bigcup_{m=1}^\infty K_m$ is also separable. We define the countable collection of closed sets

$$\mathcal{C} = \overline{\{B(x_k^{1,m}, 1/i) : 1 \leq k \leq N_i^m, i, m \in \mathbb{Z}_+\}}$$

and for each open G and each $x \in G \cap \bigcup_{m=1}^\infty K_m$ there exists a set C such that $x \in C \in \mathcal{C}$. Then, we define

$$\mathcal{H} = \left\{ \bigcup_{i=1}^n C_i \cap K_m : C_i \in \mathcal{C}, m \in \mathbb{Z}_+ \right\} \cup \{\emptyset\}$$

and consider an ordering of elements in this set $\{H_i\}_{i=1}^\infty$. Using a diagonal argument, we can find a subsequence $\{\mathbb{P}_k\}_{k=1}^\infty$ of $\{\mathbb{P}_\alpha\}_{\alpha \in \mathcal{A}}$ such that $\mathbb{P}_k(H_i) \rightarrow \beta(H_i)$ for all $H_i \in \mathcal{H}$. Now, define

$$\mathbb{P}(G) = \sup_{i: H_i \subset G} \beta(H_i),$$

and we will show that $\mathbb{P}$ is a probability measure. First

$$\mathbb{P}(S) = \sup_{i: H_i \subset S} \beta(H_i) \geq \beta(K_m) \geq 1 - 1/m.$$

Since, m is arbitrary it follows that $\mathbb{P}(S) = 1$.

Now, we will make use of outer measures to prove our final claim. First, we define

$$\gamma(M) = \inf_{G \supset M} \mathbb{P}(G).$$

If γ is an *outer measure*, then this implies that there is a σ -algebra $\mathcal{M}$ where γ as defined above is a measure on $(S, \mathcal{M})$. If then we can prove that any closed set F is in $\mathcal{M}$, this will imply that $\mathcal{B}(S) \subset \mathcal{M}$. Using the definition of γ and $\mathbb{P}(G) = \gamma(G)$ is a probability measure.

Step 1: For any closed F such that $F \subset G$ (open) and $F \subset H \in \mathcal{H}$, there is an $H_0 \in \mathcal{H}$ such that $F \subset H_0 \subset G$. $F \subset H$ implies that F is a compact set. Hence, there is a finite cover of F using the open balls used in the definition of $\mathcal{C}$, we can denote this as $B_1, B_2, \dots, B_{N_F}$. Taking the closure of these sets in finite cover and using a compact set $K_u \supset F$, we have

$$H_0 = \bigcup_{n=1}^{N_F} (\overline{B_n} \cup K_u),$$

which satisfies $F \subset H_0 \subset G$.

Step 2: $\mathbb{P}$ is finitely subadditive on the open sets. Take any $\mathcal{H} \ni H \subset G_1 \cup G_2$. We partition (with the exception of boundaries) $H = F_1 \cup F_2$, where $F_1 \subset G_1$ and $F_2 \subset G_2$ (For each $x \in H$ put x in the closed set which is farthest from the complement). Then, we can find closed sets H_i such that $F_i \subset H_i \subset G_i$, $i = 1, 2$. This implies,

$$\mathbb{P}(G_1 \cup G_2) = \sup_{n: H_n \subset G_1 \cup G_2} \beta(H_n) \leq \mathbb{P}(G_1) + \mathbb{P}(G_2).$$

Step 3: $\mathbb{P}$ is countably subadditive on the open sets. Consider $H \subset \bigcup_{i=1}^{\infty} G_i$ and compactness of H implies $H \subset \bigcup_{i=1}^N G_i$ for some finite N . Hence, $\beta(H_n) \leq \sum_{i=1}^N \mathbb{P}(G_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(G_i)$.

Step 4: γ is an outer measure, i.e., $\gamma(\emptyset) = 0$ and it is countably subadditive for arbitrary subsets of S . $\gamma(\emptyset) = 0$ is clear from the definition. Take a sequence M_n and choose open sets G_n such that $\mathbb{P}(G_n) \leq \gamma(M_n) + \epsilon/2^n$ and

$$\gamma\left(\bigcup_{n=1}^{\infty} M_n\right) \leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} G_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(G_n) \leq \sum_{n=1}^{\infty} \gamma(M_n) + \epsilon.$$

Since $\epsilon > 0$ is arbitrary the result follows.

Step 5: $\mathbb{P}(G) \geq \gamma(F \cap G) + \gamma(F^c \cap G)$ for all F closed and G open. First, we find $H_1 \in \mathcal{H}$ such that $H_1 \subset F^c \cap G$ and $\beta(H_1) > \mathbb{P}(F^c \cap G) - \epsilon$. Then choose $H_0 \in \mathcal{H}$ such that $H \in H_1^c \cap G$ and $\beta(H_1) > \mathbb{P}(H_1^c \cap G) - \epsilon$. Then, $\mathbb{P}(G) \geq \beta(H_0 \cup H_1) = \beta(H_0) + \beta(H_1) > \mathbb{P}(H_1^c \cap G) + \mathbb{P}(F^c \cap G) - 2\epsilon \geq \gamma(F \cap G) + \gamma(F^c \cap G) - 2\epsilon$ and the result follows as ϵ is arbitrary.

Step 6: Any closed $F \in \mathcal{M}$. Remember that a set F is in $\mathcal{M}$ if and only if $\gamma(L) \geq \gamma(F \cap L) + \gamma(F^c \cap L)$ for any arbitrary subset of S . For any open $G \supset L$ we have $\mathbb{P}(G) \leq \gamma(F \cap L) + \gamma(F^c \cap L)$ taking infimum over $G \supset L$ the result follows.

This ensures that $\mathbb{P}$ defined as above is the limit of the subsequence and is a probability measure.

To prove the converse, we use the fact that any closed and totally bounded set is compact. Remember that a set K is totally bounded if for every $\epsilon > 0$ it can be covered by finitely many open balls of radius ϵ , which is for separable metric spaces equivalently

$$K = \bigcap_{i=1}^{\infty} \bigcup_{k=1}^{N_j} \overline{B(x_i, 1/j)},$$

where x_i is a countable dense set in S .

If for any $j \in \mathbb{Z}_+$ we can find a unique N_j for every $\alpha \in \mathcal{A}$ such that $\mathbb{P}_{\alpha} \left(\bigcup_{i=1}^{N_j} B(x_i, 1/j) \right) > 1 - \epsilon/2^j$, using subadditivity we can prove that K is the compact set that we can use to state tightness. Suppose, this does not hold, then there exists a j_0 such that for every k we can find a $\mathbb{P}_{\alpha_k}$ such that

$$\mathbb{P}_{\alpha_k} \left(\bigcup_{i=1}^k B(x_i, 1/j_0) \right) < 1 - \epsilon/2^{j_0}.$$

The relative compactness implies that this subsequence $\{\mathbb{P}_{\alpha_k}\}_{k=1}^{\infty}$ has a convergent subsequence. However, it will also imply that $\mathbb{P}(S) < 1 - \epsilon/2^{j_0}$

□

7 Prokhorov Metric

Another way to deal with convergence in distribution (weak convergence) is to define the space of probability measures $\mathcal{M}_1$ on $(S, \mathcal{B}(S))$ and consider convergence in this space. We remind that for the space $\mathcal{M}_1$ a topology is a collection τ of subsets of $\mathcal{P}$ that satisfies the following conditions

1. $\emptyset \in \tau$ and $\mathcal{M}_1 \in \tau$
2. If $x_\alpha \in \tau$ for $\alpha \in A$ for arbitrary A , then $\bigcup_{\alpha \in A} x_\alpha \in \tau$.
3. If A is finite and $x_\alpha \in \tau$ for $\alpha \in A$, then $\bigcap_{\alpha \in A} x_\alpha \in \tau$.

For any finite collection $f_1, \dots, f_k, \epsilon > 0$ and $\mathbb{P} \in \mathcal{M}_1$ define sets

$$G = \left\{ \mathbb{Q} \in \mathcal{M}_1 : \left| \int_S f_i(x) \mathbb{Q}(dx) - \int_S f_i(x) \mathbb{P}(dx) \right| < \epsilon \right\},$$

and consider the topology τ generated by these sets. It is clear that this topology is the topology of weak convergence. It turns out that this topology is metrizable with the following metric suggested by Prohorov

$$\tilde{\rho}(\mathbb{P}, \mathbb{Q}) = \inf \{ \delta : \mathbb{P}(A) \leq \mathbb{Q}(A^\delta) + \delta \text{ and } \mathbb{Q}(A) \leq \mathbb{P}(A^\delta) + \delta, \text{ for all } A \in \mathcal{B}(S) \}.$$

We first prove that $\tilde{\rho}$ is actually a metric. It is clear that $\delta(\mathbb{P}, \mathbb{Q}) \geq 0$ and replacing A in the definition with any closed set F , $\tilde{\rho}(\mathbb{P}, \mathbb{Q}) = 0$ definition implies that $\mathbb{P}(F) = \mathbb{Q}(F)$ and hence $\mathbb{P} = \mathbb{Q}$. The symmetry is also apparent from the definition. Finally, we need to prove that triangle inequality holds, i.e., for any three probability measures $\mathbb{P}, \mathbb{Q}, \mathbb{U}$, we need to show $\tilde{\rho}(\mathbb{P}, \mathbb{U}) \leq \tilde{\rho}(\mathbb{P}, \mathbb{Q}) + \tilde{\rho}(\mathbb{Q}, \mathbb{U})$. Suppose, we have $\tilde{\rho}(\mathbb{P}, \mathbb{Q}) \leq \epsilon_1, \tilde{\rho}(\mathbb{Q}, \mathbb{U}) \leq \epsilon_2$, then

$$\mathbb{P}(A) \leq \mathbb{Q}(A^{\epsilon_1}) + \epsilon_1 \leq \mathbb{U}((A^{\epsilon_1})^{\epsilon_2}) + \epsilon_1 + \epsilon_2 \leq \mathbb{U}(A^{\epsilon_1 + \epsilon_2}) + \epsilon_1 + \epsilon_2,$$

which implies $\tilde{\rho}(\mathbb{P}, \mathbb{U}) \leq \epsilon_1 + \epsilon_2$ using symmetry and hence the triangle inequality holds.

Prokhorov's metric is related to weak convergence, but under most general conditions this is an "if" relation.

Theorem 7.1. *For a sequence of probability measures $\{\mathbb{P}_n\}_{n=1}^\infty$, if $\tilde{\rho}(\mathbb{P}_n, \mathbb{P}) \rightarrow 0$, then $\mathbb{P}_n \Rightarrow \mathbb{P}$.*

Proof. We can use Aleksandrov's theorem with closed sets to prove this claim. Using the assumption that $\tilde{\rho}(\mathbb{P}_n, \mathbb{P}) \rightarrow 0$ and a diagonal argument, for any closed set F we can find $\epsilon_n \downarrow 0$ such that $\mathbb{P}_n(F) \leq \mathbb{P}(F^{\epsilon_n}) + \epsilon_n$ and hence

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n(F) \leq \limsup_{n \rightarrow \infty} \mathbb{P}(F^{\epsilon_n}) + \epsilon_n = \mathbb{P}(F),$$

where the last step follows using continuity of $\mathbb{P}$ from above. □

If we have the additional assumption of separability of underlying space S , we can prove the implication in the reverse direction as well.

Theorem 7.2. *If the metric space S is separable and $\mathbb{P}_n \Rightarrow \mathbb{P}$, then $\tilde{\rho}(\mathbb{P}_n, \mathbb{P}) \rightarrow 0$.*

Proof. Separability ensures that for any $\epsilon > 0$, we can find a countable collection of open balls $\{A_n^\epsilon\}_{n=1}^\infty$ with diameter less than ϵ such that $\bigcup_{n=1}^\infty A_n^\epsilon = S$. Using this, we can choose an N_ϵ^1 such that $\mathbb{P}\left(\bigcup_{n=1}^\infty A_n^\epsilon\right) < \epsilon$. Using Aleksandrov's theorem, for any open set G , there is N_ϵ^2 such that $n > N_\epsilon^2$ implies $\mathbb{P}_n(G) + \epsilon > \mathbb{P}(G)$. Now, setting $N_\epsilon = \max\{N_\epsilon^1, N_\epsilon^2\}$ any Borel set can be covered by a subset of $\{A_1^\epsilon, \dots, A_{N_\epsilon}^\epsilon\} \cup \{\bigcup_{n=N_\epsilon+1}^\infty A_n^\epsilon\}$. Taking the union of all sets $\{A_1^\epsilon, \dots, A_{N_\epsilon}^\epsilon\}$ such that $A_n^\epsilon \cap G \neq \emptyset$ and calling it G_0 , we have $G_0 \subset G^\epsilon$. Hence, for all $n \geq N_\epsilon$

$$\mathbb{P}(G) \leq \mathbb{P}(G_0) + \epsilon \leq \mathbb{P}_n(G_0) + 2\epsilon \leq \mathbb{P}_n(G^{2\epsilon}) + 2\epsilon.$$

Now, if we can prove that $\mathbb{P}(A) \leq \mathbb{P}_n(A^\epsilon) + \epsilon$ for all Borel sets A imply $\mathbb{P}_n(B) \leq \mathbb{P}(B^\epsilon) + \epsilon$, again for all Borel sets B the result follows:

$$\begin{aligned} \mathbb{P}(B^\epsilon) &= 1 - \mathbb{P}((B^\epsilon)^c) \\ &\geq 1 - \mathbb{P}_n(((B^\epsilon)^c)^\epsilon) - \epsilon \\ &\geq \mathbb{P}_n(B) - \epsilon. \end{aligned}$$

The last equality follows as $B \subset (((B^\epsilon)^c)^\epsilon)^c$. $\square$

Using the Prokhorov metric, we can treat the set of probability measures on a separable metric space S as a metric space where convergence means weak convergence. We can also ask whether this new space is separable and/or complete. Fortunately, the space of probability measures inherits these properties from the underlying metric space S as proven in the following theorem:

Theorem 7.3. *If the metric space S is separable and complete, then the space of probability measures on $(S, \mathcal{B}(S))$ equipped with Prokhorov metric is also separable and complete.*

Proof. The proof of separability is similar to the proof of Theorem 7.2. For any probability measure $\mathbb{P}$, again choose an N_ϵ^1 such that $\mathbb{P}\left(\bigcup_{n=1}^\infty A_n^\epsilon\right) < \epsilon$ and choose an $x_i \in A_i$ for all $i = 1, \dots, N_\epsilon^1$ and $x_{N_\epsilon^1+1} \in \bigcup_{n=N_\epsilon^1+1}^\infty A_n^\epsilon$. Now, consider the discrete probability measure $\mathbb{P}_d(\{x_i\}) = r_i$ for each $i = 1, \dots, N_\epsilon^1 + 1$ such that

- (i) $\sum_{i=1}^{N_\epsilon^1+1} r_i = 1$
- (ii) $r_{N_\epsilon^1+1} = \epsilon$
- (iii) $\sum_{i=1}^{N_\epsilon^1} |r_i - \mathbb{P}(A_i)| < \epsilon$.

Using the same argument in the proof of Theorem 7.2, we can show that for all Borel sets A ,

$$\mathbb{P}(A) \leq \mathbb{P}_d(A) + 2\epsilon,$$

again using the last part of the above proof, this shows that for any probability measure on $(S, \mathcal{B}(S))$ we can find a discrete probability measure with finite support arbitrarily close. Hence, the space of probability measures is separable.

To prove that $\mathcal{P}$ is complete, we need to analyze the fundamental sequences $\{\mathbb{P}_n\}_{n=1}^\infty$, where for any $\epsilon > 0$ there exists N_ϵ such that $k, l > N_\epsilon$ implies $\hat{\rho}(\mathbb{P}_k, \mathbb{P}_l) < \epsilon$. Take $N_{\epsilon/2}$ for $\epsilon/2$ and choose the compact K such that $\sup_{1 \leq n \leq N_{\epsilon/2}} \mathbb{P}_n(K) > 1 - \epsilon/2$. This proves that $\{\mathbb{P}_n\}_{n=1}^\infty$ is tight and have a subsequence that converges. Then as in the proof of Prokhorov's Theorem, we can construct the limiting probability distribution. $\square$

8 Some Basic Properties of Weak Convergence

8.1 Skorokhod's Representation Theorem

Consider the sequence of random elements $\{X_n\}_{n=1}^{\infty}$ with the corresponding probability measures $\{\mathbb{P}_n\}_{n=1}^{\infty}$. If $\mathbb{P}_n \Rightarrow \mathbb{P}$ and X has distribution $\mathbb{P}$ we also say X_n converges weakly (or in distribution) to X . However, this mode of convergence does not necessarily have any implications on the actual values of the random variables. Here is a very basic example:

Example 8.1. Consider i.i.d. random variables $X, X_1, X_2, \dots$ with $N(0, 1)$ distribution. It is clear that $X_n \Rightarrow X$, but using Borel-Cantelli it is easy to prove that the observations does not converge with probability 1.

In fact, these variables may not even be defined on the same probability space! This causes problems when we need to conclude about the actual values. Skorokhod's representation theorem says that even though the values of random variables at hand does not converge, we can always find a space where we can define random variables $\tilde{X}_n, \tilde{X}$ with the same distribution as X_n, X almost surely.

Theorem 8.2. Suppose $\mathbb{P}_n \Rightarrow \mathbb{P}$, on a separable measurable metric space $(S, \mathcal{B}(S))$. Then there exists a common space $(\Omega, \mathcal{F})$ and random elements X_n with probability distributions $\mathbb{P}_n$ and X with $\mathbb{P}$ such that $X_n \rightarrow X$ almost everywhere.

Proof. The assumed weak convergence assumes tightness and hence for any $\epsilon > 0$, we can find a compact K such that

$$\sup_n \mathbb{P}_n[K^c] < \epsilon \text{ and hence, } \mathbb{P}(K^c) < \epsilon.$$

Also, using compactness it is possible to find a finite partition $B_0, B_1, B_2, \dots, B_N$ of S such that each set is contained in a ball of radius ϵ and $\mathbb{P}(\partial B_i) = 0$ for all $i = 1, \dots, n$. Considering $B_0 = K^c \cup \bigcup_{i: \mathbb{P}(B_i)=0} B_i$ we can assume that all $\mathbb{P}(B_i) > 0$, for $i = 1, \dots, n$. Suppose that we choose for any $\epsilon = 1/2^m$, we choose sets $B_0^m, B_1^m, \dots, B_{N_m}^m$ to satisfy these properties. For any $m \geq 0$, we can find an n_m such that

$$\mathbb{P}_n(B_k^m) \geq (1 - \epsilon_m) \mathbb{P}(B_k^m) \text{ for all } k.$$

We see that any given n there exists an m such that $n_m \leq n < n_{m+1}$. Now the idea is to find a "coupling" of $\tilde{X}_n$ and $\tilde{X}$ for all n such that

- (i) $\tilde{X}_n \sim X_n$ for all n and $\tilde{X} \sim X$.
- (ii) Suppose $n_m \leq n < n_{m+1}$, then given $\tilde{X} \in B_k^m$ then $\tilde{X}_n$ will be in B_k^m with probability $1 - \epsilon_m$, or with probability $1 - \epsilon$ it can be anywhere according to a probability distribution that makes (i) true (use the law of total probability).

To do so, we define a series of independent random elements $\tilde{X}, Y_{nk}, Z_n$, with the following distributions:

$$\begin{aligned} \mathbb{P}_{\tilde{X}}(A) &= \mathbb{P}(A) \\ \mathbb{P}_{Y_{nk}}(A) &= \frac{\mathbb{P}_n(A \cup B_k^m)}{\mathbb{P}_n(A \cup B_k^m)}, \text{ for all } m, k \text{ and } n_m \leq n < n_{m+1} \\ \mathbb{P}_{Z_n}(A) &= \epsilon_m^{-1} \sum_{k=1}^{N_m} \mathbb{P}_n(A|B_k^m) (\mathbb{P}_n(B_k^m) - (1 - \epsilon_m) \mathbb{P}(B_k^m)). \end{aligned}$$

We now define $\tilde{X}_n$ such that if $\tilde{X} \in B_k^m$ then $\tilde{X}_n \in B_k^m$ with probability $1 - \epsilon_m$ by making it equal to Y_{km} or it will be anywhere with probability ϵ_m by being equal to Z_n . This is achieved by generating a uniform(0,1) random variable ξ independent of all other parameters and setting

$$\tilde{X}_n = \sum_{k=1}^{N_m} I(\xi \leq (1 - \epsilon_m), \tilde{X} \in B_k^m) Y_{nk} + 1(\xi > (1 - \epsilon_m)) Z_n.$$

Now, define $E_m = \{X \in B_0^m, \xi < 1 - \epsilon_m\}$, we see that $\mathbb{P}(E_m) \rightarrow 0$ using Borel-Cantelli lemma, which proves that $\tilde{X}_n \rightarrow \tilde{X}$ for all ω . $\square$

The “coupling” technique used in the proof of the theorem. This is a very useful tool in probability and we refer the interested reader to Thorisson [5].

8.2 Continuous Mapping Theorem

Now, we see the Skorokhod representation theorem in action. Suppose we know that $X_n \Rightarrow X$. Then, what can be said about convergence of $g(X_n)$? Define D_g as the set of discontinuity points of $g(\cdot)$, then we have the following theorem:

Theorem 8.3. *If $\mathbb{P}_X(D_g) = 0$, then $g(X_n) \Rightarrow g(X)$.*

Proof. We concentrate on $\tilde{X}_n, \tilde{X}$ as defined in the proof of representation theorem. Then, $g(\tilde{X}_n) \rightarrow g(\tilde{X})$ when $\tilde{X}$ assumes a continuity point of $g(\cdot)$. Using the fact that $g(\tilde{X}_n) \sim g(X_n)$ and $g(\tilde{X}) \sim g(X)$ the result follows. $\square$

References

- [1] P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1997.
- [2] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, California, 1995.
- [3] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, 2002.
- [4] W. Rudin. *Functional Analysis*. McGraw-Hill, Boston, 1991.
- [5] H. Thorisson. *Coupling, Stationary, and Regeneration*. Springer-Verlag, New York, 2000.