NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

MASTER THESIS

Weak Convergence of Probability Measures and Empirical Process Theory

---

Author:
Georgios GAVRILOPOULOS

Supervisor:
Ass. Prof. Dr. Samis TREVEZAS

Department of Mathematics

October 2024

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Department of Mathematics

Master In Pure Mathematics

Weak Convergence of Probability Measures and Empirical Process Theory

by Georgios GAVRILOPOULOS

This thesis presents the fundamental principles of Weak Convergence of Probability Measures and Empirical Process Theory.

In the first part of the thesis, we introduce weak convergence and explain its relationship with the classical concept of convergence in distribution. We present the proof and some applications of the landmark theorem of Prohorov and then we focus on spaces of continuous functions and on Donsker's theorem. Finally, we discuss measurability issues that arise in the study of weak convergence and we give an overview of the potential remedies that have been suggested in the literature.

The second part of the thesis focuses on the universal law of large numbers (ULLN). We start with the classical result of Glivenko-Cantelli and we continue with generalizations of this result to more complex function spaces. We investigate the deeper connection between the ULLN and the complexity of the corresponding function classes using complexity measures such as the entropy and the Rademacher complexity. Finally, we discuss the importance of the ULLN for statistics and machine learning and we illustrate how empirical process theory can be used to derive convergence rates for parametric and non-parametric least squares estimators.

Περίληψη

Ασθενής Σύγκλιση Μέτρων Πιθανότητας και Θεωρία Εμπειρικών Διαδικασιών

Η παρούσα διπλωματική εργασία παρουσιάζει τις θεμελιώδεις αρχές της Ασθενούς Σύγκλισης Μέτρων Πιθανότητας και της Θεωρίας Εμπειρικών διαδικασιών.

Στο πρώτο μέρος της εργασίας κάνουμε μια εισαγωγή στη θεωρία της ασθενούς σύγκλισης και εξηγούμε τη σχέση της με την κλασική έννοια της σύγκλισης κατά κατανομή. Παρουσιάζουμε την απόδειξη και κάποιες εφαρμογές του θεωρήματος-ορόσημου του Prohorov και στη συνέχεια εστιάζουμε σε χώρους συνεχών συναρτήσεων και στο θεώρημα του Donsker. Τέλος, συζητούμε κάποια ζητήματα μετρησιμότητας που εμφανίζονται στην μελέτη της ασθενούς σύγκλισης, και κάνουμε μια επισκόπηση των πιθανών τρόπων αντιμετώπισης που έχουν προταθεί στη βιβλιογραφία.

Το δεύτερο μέρος της εργασίας εστιάζει στον ομοιόμορφο νόμο των μεγάλων αριθμών. Ξεκινούμε τη μελέτη με το κλασικό θεώρημα των Glivenko-Cantelli και συνεχίζουμε με γενικεύσεις αυτού του θεωρήματος σε γενικότερους χώρους συναρτήσεων. Ερευνούμε τη βαθύτερη σύνδεση του ομοιόμορφου νόμου των μεγάλων αριθμών με την πολυπλοκότητα των αντίστοιχων συναρτησιακών χώρων μέσω μέτρων πολυπλοκότητας όπως η εντροπία και η πολυπλοκότητα Rademacher. Τέλος, συζητούμε τη σημασία του ομοιόμορφου νόμου των μεγάλων αριθμών για τη στατιστική και τη μηχανική μάθηση και παρουσιάζουμε τρόπους με τους οποίους μπορούμε να χρησιμοποιήσουμε τη θεωρία εμπειρικών διαδικασιών για να βγάλουμε συμπεράσματα για το ρυθμό σύγκλισης παραμετρικών και μη-παραμετρικών εκτιμητών ελαχίστων τετραγώνων.

Ευχαριστίες

Με την ολοκλήρωση αυτού του μεταπτυχιακού προγράμματος, θα ήθελα πρωτίστως να ευχαριστήσω τον καθηγητή μου, Σάμη Τρέβεζα, για την απεριόριστη υπομονή του, καθώς και την συνεχή στήριξη που μου παρείχε. Χωρίς αυτή την πολύτιμη στήριξη, δε θα είχα καταφέρει να φτάσω μέχρι το τέλος του δρόμου. Ευχαριστώ επίσης την οικογένειά μου, και ιδιαιτέρως τον πατέρα μου, Παναγιώτη, για τις εποικοδομητικές παρεμβάσεις του κατά τη συγγραφή της διπλωματικής μου εργασίας.

Contents

Chapter 1

Weak Convergence of Probability Measures

In the first part of this thesis, we aim to give a comprehensive introduction to the theory of weak convergence of probability measures. Before moving on to the formal exposition of this theory, we shall present some elements about its early stages and its development through time. In this preliminary discussion, we shall also attempt to explain why the theory of weak convergence is important from a statistical point of view.

First, we should give some elementary definitions. We begin with the definition of the *empirical distribution function.*

Definition 1.0.1. Let n be a positive integer and let $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ be a random sample from a distribution with cumulative distribution function (c.d.f.) F. For all $x \in \mathbb{R}$ we define the random variable $\mathbb{F}_n(x)$ by

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}},$$

where $\mathbf{1}_A$ denotes the indicator function of a set A. The function $\mathbb{F}_n$ is called the empirical distribution function of the sample $(X_1, X_2, \ldots, X_n)$.

Sometimes, the term *empirical distribution function* will be used in a slightly different context. More specifically, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, then $\mathbb{F}_n(x, \cdot)$ is the random variable [1] defined by

$$\mathbb{F}_n(x, \omega) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i(\omega) \leq x\}}, \ \omega \in \Omega.$$

[1] Notice that, since $X_1, X_2, \ldots, X_n$ are random variables (and thus measurable with respect to the $\sigma-$fields of Ω and $\mathbb{R}$), the quantity $\mathbb{F}_n(x, \cdot)$ is also measurable. Hence, it makes sense to refer to $\mathbb{F}_n(x, \cdot)$ with the term *random variable*

In this case, the empirical distribution function of $(X_1, \ldots, X_n)$ is the stochastic process $\{\mathbb{F}_n(x, \cdot)\}_{x \in \mathbb{R}}$.

Let x be an arbitrary real number. From the strong law of large numbers and the fact that $\mathbf{1}_{\{X_i(\omega) \leq x\}}$ is a Bernoulli random variable with success probability equal to $F(x)$, it follows immediately that $\mathbb{F}_n(x, \cdot) \to F(x)$ almost surely (a.s.). As we shall see in the next chapter, $\mathbb{F}_n(x, \cdot)$ is "close" to $F(x)$ in a stronger sense.

If $(\Omega, \mathcal{A}, P)$ is a probability space and $(T, \mathcal{T})$ is a measurable space, then an $\mathcal{A}/\mathcal{T}-$ measurable function $X : \Omega \to T$ is called a *random element* of T. A random variable is simply a special case of a random element, in which $T = \mathbb{R}$ and $\mathcal{T} = \mathcal{B}(\mathbb{R})$.

If $C = C[0, 1]$ is the space of continuous functions $f : [0, 1] \to \mathbb{R}$ equipped with the supremum norm and the corresponding Borel σ-field $\mathcal{C}$, then a random element X of C is simply a stochastic process with continuous sample paths. Such a process is called *Gaussian* if

$$(X_{t_1}, \ldots, X_{t_k})$$

is normally distributed for all $t_1, \ldots, t_k \in [0, 1]$. The *Wiener process* is a Gaussian random element $\{\mathbb{W}_t\}_{t \in [0,1]}$ of C satisfying the following conditions:

- $\mathbb{W}_0 \equiv 0$,
- If $0 = t_0 \leq t_1 \leq \ldots \leq t_k = 1$, then the random variables

$$\mathbb{W}_{t_1} - \mathbb{W}_{t_0}, \ \ldots, \ \mathbb{W}_{t_k} - \mathbb{W}_{t_{k-1}}$$

  are independent,
- For all $t \in [0, 1]$ it holds that $\mathbb{W}_t \sim \mathcal{N}(0, t)$.

A *Brownian bridge* is a random element $\mathbb{B}$ of C defined by

$$\mathbb{B}_t = \mathbb{W}_t - t\mathbb{W}_1,$$

where $\mathbb{W}$ is a Wiener process.

The starting point of the theory of weak convergence can be considered to be the problem of *goodness of fit*. In this problem, we are given n independent observations from a distribution with c.d.f. F, and we want to test the null hypothesis

$$H_0 : \ F = F_0 \text{ vs } H_1 : F \neq F_0,$$

where F_0 is a fixed distribution function. To deal with this problem, A. Kolmogorov used the statistic

$$D_n := \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F_0(x)| .$$

As it was mentioned earlier, $\mathbb{F}_n(x)$ should be "close" to the underlying distribution function, so the null hypothesis is rejected for "large" values of the statistic.

In order to construct a statistical test based on D_n we need to determine the distribution of D_n. An interesting result is that the asymptotic distribution of $K_n = \sqrt{n} D_n$ does not depend on the underlying c.d.f. F. Thus, we can determine it using uniformly distributed random variables. In other words, that distribution is the same as the asymptotic distribution of

$$\sup_{t \in [0,1]} |\mathbb{U}_n(t)| ,$$

where

$$\mathbb{U}_n(t) := \sqrt{n} \left(\mathbb{F}_n(t) - t \right) ,$$

and $\mathbb{F}_n$ is the empirical c.d.f. constructed from a sequence $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$. The process $\{\mathbb{U}_n(t, \cdot)\}_{t \in [0,1]}$ is referred to as the *uniform empirical process*. In 1933, Kolmogorov [Kolmogorov, 1933] presented a thorough description of the asymptotic distribution of D_n using elementary methods.

In 1949, J. Doob [Doob, 1949] treated

$$K_n = \sqrt{n}\, D_n = \sqrt{n} \sup_{t \in [0,1]} |\mathbb{F}_n(t) - t|$$

as a function of the uniform empirical process $\mathbb{U}_n(t)$. In other words, he observed that $K_n = f(\mathbb{U}_n)$, where

$$f(x) = \sup_{t \in [0,1]} |x(t)| , \text{ for } x : [0,1] \to \mathbb{R}.$$

Doob subsequently looked for a general approach to obtain the asymptotic distribution of such quantities. Using the multivariate Central Limit Theorem and the fact that the random variable $n\mathbb{F}_n(t),\ t \in [0,1]$ has a binomial distribution with probability of success equal to t, he deduced that, for any $t_1, \ldots, t_k$, the random vector

$$(\mathbb{U}_n(t_1), \ldots, \mathbb{U}_n(t_k))$$

converges in distribution to $\mathcal{N}_k(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \min(t_i, t_j) - t_i t_j$. This distribution is the same as that of the random vector

$$(\mathbb{U}_{t_1}, \ldots, \mathbb{U}_{t_k}),$$

where $\{\mathbb{U}(t)\}_{t\in[0,1]}$ is the Brownian bridge. Thus, Doob conjectured that $\mathbb{U}_n$ must converge in some sense to $\mathbb{U}$. Hopefully, this, yet unknown, notion of convergence, would allow us to deduce that $f(\mathbb{U}_n)$ converges in distribution to $f(\mathbb{U})$.

In 1951, M. Donsker [Donsker, 1951] developed a new notion of convergence of stochastic processes, the so-called *weak convergence*. In 1952, he showed that the uniform empirical process $\mathbb{U}_n$ converges weakly to the Brownian process $\mathbb{U}$ [Donsker, 1952]. Moreover, Donsker's results showed that several functionals of $\mathbb{U}_n$ converge in distribution to the corresponding functionals of $\mathbb{U}$. In particular, it yields that

$$\sup_{t\in[0,1]} |\mathbb{U}_n(t)| \xrightarrow{d} \sup_{t\in[0,1]} |\mathbb{U}_t|,$$

which essentially leads to an accurate description of the asymptotic distribution of D_n.

During the 1950s, the works of Y. Prohorov [Prokhorov, 1956] and A. V. Skorokhod [Skorokhod, 1956] led to the major development of Donsker's theory of weak convergence of stochastic processes. Finally, in 1968, P. Billingsley [Billingsley, 1968] summarized the above works and presented a more general theory of convergence of probability measures. In what follows, we present the most significant results of this theory as well as some of its extensions.

1.1 The Classic Theory of Weak Convergence

1.1.1 Definitions and First Results

Let F and $\{F_n\}_{n=1}^{\infty}$ be arbitrary cumulative distribution functions (c.d.f.) defined on the real line. We say that F_n *converges in distribution (or converges weakly)* to F if

$$F_n(x) \to F(x), \quad \text{as } n \to \infty, \tag{1.1}$$

for all continuity points x of F. We indicate weak convergence by writing $F_n \xrightarrow{d} F$.

Let P and $\{P_n\}_{n=1}^{\infty}$ be the Borel probability measures generated by F and the sequence F_n respectively. This means that P and P_n are uniquely determined by the relations

$$P_n(-\infty, x] = F_n(x) \text{ and } P(-\infty, x] = F(x)$$

for all $x \in \mathbb{R}$ and $n \in \mathbb{N}$. We know that F is continuous at a point x if and only if $P\{x\} = 0$, so $F_n \xrightarrow{d} F$ means that the implication

$$P_n(-\infty, x] \to P(-\infty, x], \text{ if } P\{x\} = 0$$

is true for all $x \in \mathbb{R}$. Using the notation ∂A for the boundary of the set $A = (-\infty, x]$, the above relation is written as

$$P_n(A) \to P(A), \text{ if } P(\partial A) = 0. \tag{1.2}$$

In this chapter we shall prove that $F_n \xrightarrow{d} F$ if and only if the implication (1.2) is true *for any* Borel set A. Any Borel set A with $P(\partial A) = 0$ will be called a $P-$*continuity set*. We will say that $P_n \Rightarrow P$ if $P_n(A) \to P(A)$ is true for all $P-$continuity sets A, that is, if (1.2) holds. In other words, $P_n \Rightarrow P$ if and only if the corresponding distribution functions F_n and F satisfy $F_n \xrightarrow{d} F$.

The main purpose of this section is the extension of the notion of weak convergence from $\mathbb{R}$ to general metric spaces. More specifically, let T be an arbitrary metric space, and let $\mathcal{B}(T)$ be its Borel $\sigma-$field. If P_n and P are probability measures defined on $\mathcal{B}(T)$, we shall give a definition of the weak convergence $P_n \Rightarrow P$ and show that it indeed generalizes the familiar concept of weak convergence discussed above.

For any probability measure Q defined on $\mathcal{B}(T)$ and any Borel measurable function $f : T \to \mathbb{R}$ we denote

$$Qf := \int_T f \, dQ.$$

Also, a probability measure defined on the Borel $\sigma-$field $\mathcal{B}(T)$ will be called a *Borel probability measure*.

Definition 1.1.1. Let T be a metric space, and let $\mathcal{B}(T)$ be its Borel $\sigma-$field. Suppose that $\{P_n\}_{n=1}^{\infty}$ and P are probability measures defined on $\mathcal{B}(T)$. We say that P_n *converges weakly* to P, and we write $P_n \Rightarrow P$ if

$$P_n f \to Pf \tag{1.3}$$

for all bounded, continuous functions $f : T \to \mathbb{R}$.

Recall that, if T is a metric space and $\mathcal{T}$ is a $\sigma-$field on T, then a measure μ on $\mathcal{T}$ is called *regular* if, for any $A \in \mathcal{T}$ we have

- $\mu(A) = \sup\{\mu(F) : \ F \in \mathcal{T}, \ F \text{ closed}\}$,
- $\mu(A) = \inf\{\mu(G) : \ G \in \mathcal{T}, \ G \text{ open}\}$.

If P is a Borel probability measure on T, then it is easy to see that P is regular if and only if, for any set $A \in \mathcal{B}(T)$ and any $\varepsilon > 0$, there exist an open set G and a closed set F such that $F \subset A \subset G$ and $P(G \backslash F) < \varepsilon$.

Theorem 1.1.2. Every Borel probability measure P on $(T, \mathcal{T})$ is regular.

Proof. Let $\mathcal{G}$ be the class of sets $A \subset T$ such that, for all $\varepsilon > 0$, there exist sets G and F as in the above paragraph. We shall show that $\mathcal{G}$ is a $\sigma-$field that contains the closed sets, which will give us that $\mathcal{B}(T) \subset \mathcal{G}$ (which is exactly what we want to show). Let d be the metric on T and let $dist(x, A)$ be the distance of an element $x \in T$ from a set $A \subset T$ [2].

We shall first show that $\mathcal{G}$ contains the closed sets. Let A be a closed set. We take $F = A$. The sets $G_n = A^{1/n} = \left\{x \in T : d(x, A) < \frac{1}{n}\right\}$, $n = 1, 2, \ldots$ are open because they are the inverse images of the open sets $\left(-\infty, \frac{1}{n}\right)$ with respect to the continuous function $x \mapsto dist(x, A)$[3]. Also,

$$\bigcap_{n=1}^{\infty} G_n = \{x \in T : d(x, A) = 0\} = \overline{A} = A$$

so, since the sequence G_n is decreasing, it follows from the continuity of the measure that

$$\lim_{n \to \infty} P(G_n) = P(A).$$

Thus, for any $\varepsilon > 0$, we can find $n \in \mathbb{N}$ such that $0 \leq P(G_n) - P(A) < \varepsilon$. We take $G = G_n$. Since $F = A$ and $A \subset G_n$, the latter relation can be written as $P(G \backslash F) < \varepsilon$. It follows that $A \in \mathcal{G}$, so $\mathcal{G}$ contains all the closed sets in T.

We are now going to show that $\mathcal{G}$ is a $\sigma-$field. It is obvious that $\mathcal{G}$ is nonempty, and it is easy to see that it is closed under complements. Indeed, if $A \in \mathcal{G}$ and $\varepsilon > 0$, we can find sets G and F as above. Then, F^c is open, G^c is closed, $G^c \subset A^c \subset F^c$, and $P(F^c \backslash G^c) = P(F^c) - P(G^c) = (1 - P(F)) - (1 - P(G)) = P(G) - P(F) = P(G \backslash F) < \varepsilon$, so $A^c \in \mathcal{G}$. Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of sets in $\mathcal{G}$ and let ε be a positive real number. For each A_n, we can find an open set G_n and a closed set F_n such that $F_n \subset A_n \subset G_n$ and $P(G_n \backslash F_n) < \varepsilon / 2^{n+1}$. We consider a positive integer n_0 such that

$$P\left(\bigcup_{n=n_0+1}^{\infty} F_n\right) < \frac{\varepsilon}{2}.$$

[2] $dist(x, A) = \inf\{d(x, y) : y \in A\}$

[3] This function is Lipschitz continuous with Lipschitz constant equal to 1.

Such a positive integer exists by the continuity of the measure, because the sequence $B_n = \bigcup_{k=1}^{n} F_k$ is increasing and its union is $\bigcup_{k=1}^{\infty} F_k$ [4]. We now consider the sets

$$G = \bigcup_{n=1}^{\infty} G_n \text{ and } F = \bigcup_{n=1}^{n_0} F_n.$$

The first one is open and the second is closed, and it is true that $F \subset \bigcup_{n=1}^{\infty} A_n \subset G$. Also, since $(\bigcup_{n=1}^{\infty} G_n \backslash \bigcup_{n=1}^{\infty} F_n) \subset \bigcup_{n=1}^{\infty} (G_n \backslash F_n)$, it follows that

$$\begin{aligned} P(G\backslash F) &= P(G) - P(F) \\ &= P(G) - P\left(\bigcup_{n=1}^{\infty} F_n\right) + P\left(\bigcup_{n=1}^{\infty} F_n\right) - P(F) \\ &< P\left(\bigcup_{n=1}^{\infty} G_n \backslash \bigcup_{n=1}^{\infty} F_n\right) + \frac{\varepsilon}{2} \\ &\leq P\left(\bigcup_{n=1}^{\infty} (G_n \backslash F_n)\right) + \frac{\varepsilon}{2} \\ &< \left(\sum_{n=1}^{\infty} \frac{\varepsilon}{2^{n+1}}\right) + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

and the proof is complete. □

The above theorem shows that a Borel probability measure is uniquely determined by the values $P(F)$ for closed sets F. Indeed, given these values, we can determine the values $P(G)$ for all open sets G. Then, using regularity, we can find $P(A)$ for all Borel sets A. The next theorem implies that P is also determined by the values Pf for certain functions $f : T \to \mathbb{R}$.

Theorem 1.1.3. Two probability measures P, Q on $\mathcal{B}(T)$ are equal if $Pf = Qf$ for all bounded, uniformly continuous functions $f : T \to \mathbb{R}$.

Proof. Let F be a closed set and let ε be a positive real number. We consider the function $f : T \to \mathbb{R}$ with

$$f(x) = \left(1 - \frac{dist(x,F)}{\varepsilon}\right)^{+} = \begin{cases} 1 - (dist(x,F)/\varepsilon) & \text{if } dist(x,F) < \varepsilon \\ 0 & \text{else.} \end{cases}$$

It is obvious that this function is bounded. We will show that $|f(a) - f(b)| \leq d(a,b)/\varepsilon$ for all $a, b \in T$. If $f(a) = f(b) = 0$, this is obviously true. If $f(a), f(b) \neq 0$, we can show the above inequality using the fact that $x \mapsto dist(x,F)$ is Lipschitz continuous with Lipschitz constant equal to 1. Assume that $f(b) = 0$ and $f(a) \neq 0$ (the other case is treated similarly). Then, $|f(a) - f(b)| = 1 - (dist(a,F)/\varepsilon)$,

[4] Therefore, there exists a integer $n_0 \geq 1$ such that $P\left(\bigcup_{k=n_0+1}^{\infty} F_k\right) = P\left(\bigcup_{k=1}^{\infty} F_k\right) - P\left(B_{n_0}\right) < \varepsilon/2$.

so we need to show that $d(a,b) \geq \varepsilon - dist(a,F)$. However, from $f(b) = 0$ we get that $dist(b,F) \geq \varepsilon$, so the Lipschitz continuity of the distance function implies that

$$\varepsilon - dist(a,F) \leq dist(b,F) - dist(a,F) \leq d(a,b).$$

It follows that f is bounded and uniformly continuous.

We are now going to show that

$$\mathbf{1}_F(x) \leq f(x) \leq \mathbf{1}_{F^\varepsilon}(x), \tag{1.4}$$

where $\mathbf{1}_A$ denotes the indicator function of the set A, and

$$F^\varepsilon = \{x \in T : dist(x,F) < \varepsilon\}.$$

If $\mathbf{1}_F(x) = 1$, then $x \in F$ so $f(x) = 1$. The left side of the inequality follows from this observation. As for the right side, we observe that $f(x) \leq 1$ for all $x \in T$. Also, if $\mathbf{1}_{F^\varepsilon}(x) = 0$, then $x \notin F^\varepsilon$ so $dist(x,F) \geq \varepsilon$, which implies that $f(x) = 0$. The right inequality follows. Integrating all three sides of the inequality with respect to P and Q we obtain

$$P(F) \leq Pf \leq P(F^\varepsilon) \text{ and } Q(F) \leq Qf \leq Q(F^\varepsilon).$$

Since $Pf = Qf$, it follows that $P(F) \leq Q(F^\varepsilon)$.

Just like in the proof of Theorem 1.1.2, we can use the continuity of the measure to show that $Q(F^\varepsilon) \to Q(F)$ as $\varepsilon \searrow 0$. Thus, $P(F) \leq Q(F)$. The converse inequality follows in a similar way. We finally get that $P(F) = Q(F)$ for all closed sets $F \subset T$, so the two measures are equal. □

The following theorem, known as the *Portmanteau Theorem* shows that weak convergence, in the way we defined it earlier, indeed extends the concept of convergence in distribution discussed at the start of the section.

Theorem 1.1.4 (Portmanteau Theorem)**.** Let T be a metric space and let $\{P_n\}_{n=1}^\infty$ and P be Borel probability measures on T. The following conditions are equivalent:

(i) $P_n \Rightarrow P$

(ii) $P_n f \to Pf$ for all bounded, uniformly continuous functions $f : T \to \mathbb{R}$.

(iii) $\limsup_n P_n(F) \leq P(F)$ for all closed sets $F \subset T$.

(iv) $\liminf_n P_n(G) \geq P(G)$ for all open sets $G \subset T$.

(v) $P_n(A) \to P(A)$ for all $P-$continuity sets $A \subset T$.

Proof. The implication (i) $\Rightarrow$ (ii) follows directly from definition 1.1.1.

For (ii) $\Rightarrow$ (iii) we shall use the function f that was used in the proof of the theorem 1.1.3. This function is bounded and uniformly continuous, so (1.4) (integrated with respect to P_n) implies that $P_n(\chi_F) \leq P_n f$, which is equivalent to $P_n(F) \leq P_n f$. It follows that

$$\limsup_n P_n(F) \leq \limsup_n P_n f = Pf.$$

However, the same inequality (integrated with respect to P) gives $Pf \leq P(F^\varepsilon)$. It follows that $\limsup_n P_n(F) \leq P(F^\varepsilon)$. Letting $\varepsilon \searrow 0$, and using again the fact that $P(F^\varepsilon) \to P(F)$, we get that $\limsup_n P_n(F) \leq P(F)$.

The implication (iii) $\Rightarrow$ (iv) follows easily by taking complements.

We will now show that (iii) and (iv) together imply (v). Conditions (iii) and (iv) together with the obvious inequalities $\overline{A} \supset A \supset A^\circ$, imply that

$$P(\overline{A}) \geq \limsup_n P_n(\overline{A}) \geq \limsup_n P_n(A) \geq \liminf_n P_n(A) \geq \liminf_n P_n(A^\circ) \geq P(A^\circ).$$

If A is a $P-$continuity set, then $P(\overline{A}) - P(A^\circ) = P(\overline{A} \backslash A^\circ) = P(\partial A) = 0$, so the first and the last terms in the above inequality are equal. Moreover, since $\overline{A} \supset A \supset A^\circ$, these terms are equal to $P(A)$ This shows that all intermediate terms are equal to $P(A)$, so

$$\limsup_n P_n(A) = \liminf_n P_n(A) = P(A),$$

whence it follows that $P_n(A) \to P(A)$.

For (v) $\Rightarrow$ (i), we will use the equality

$$Pf = \int_0^\infty P\{f > t\}\, dt,$$

which follows from Fubini's theorem [5]. We need to show that $P_n f \to Pf$ for all bounded, continuous $f : T \to \mathbb{R}$. Due to linearity and the fact that f is bounded, we may assume that $0 < f < 1$. Then, $Pf = \int_0^\infty P(\{f > t\})\, dt = \int_0^1 P(\{f > t\})\, dt$, and $P_n f = \int_0^\infty P_n(\{f > t\})\, dt = \int_0^1 P_n(\{f > t\})\, dt$. Since

[5] $\int_T f(x)\, dP(x) = \int_T \int_0^\infty \chi_{[0,f(x)]}(t)\, dt\, dP = \int_0^\infty \int_T \chi_{[0,f(x)]}(t)\, dP\, dt = \int_0^\infty \int_T \chi_{\{f(x) \geq t\}}(x)\, dP\, dt$

f is continuous, $\partial\{f > t\} = \{f = t\}$. The sets $\{f = t\}$, $t \in [0,1]$ form a partition of T, so at most countably many of them can have positive measure (because T has finite measure) [6] Thus, $\{f > t\}$ is a $P-$continuity set except for countably many $t \in [0,1]$, which gives that $P_n(\{f > t\}) \to P(\{f > t\})$ λ-almost surely in $[0,1]$, where λ is the Lebesgue measure on $[0,1]$ restricted on the corresponding Borel sets. Thus, the bounded convergence theorem gives

$$P_n f = \int_0^1 P_n(\{f > t)\})\, dt \to \int_0^1 P(\{f > t\})\, dt = Pf.$$

□

Another useful criterion for weak convergence is the following one:

Theorem 1.1.5. The following are equivalent:

(i) $P_n \Rightarrow P$.

(ii) Each subsequence $\{P_{n_i}\}$ contains a further subsequence $\{P_{n_{i_m}}\}$ such that $P_{n_{i_m}} \Rightarrow P$ as $m \to \infty$.

Proof. The direction (i) $\Rightarrow$ (ii) is obvious. For the converse, suppose that $P_n \not\Rightarrow P$. Then, $P_n f \not\to Pf$ for some bounded continuous real function f. It follows that there exists some positive $\varepsilon > 0$ and some subsequence P_{n_i} such that $|P_{n_i} f - Pf| > \varepsilon$ for all i. This subsequence does not have a subsequence that converges weakly to P, contradiction. □

As we discussed in section 1, the convergence of a sequence of stochastic processes to a limit process should imply the convergence in distribution of some function of these stochastic processes to the corresponding function of the limit process. One of the most important results of the theory of weak convergence, which ensures this property for various choices of the function, is the *Continuous Mapping Theorem*.

Before stating this result, let us first consider metric spaces $(S, \mathcal{B}(S)), (T, \mathcal{B}(T))$ together with their Borel $\sigma-$fields, and a Borel measurable function $h : S \to T$, whose set of discontinuity points is denoted by D_h. If P is a Borel probability measure defined on $\mathcal{B}(S)$, then h induces a Borel probability measure Ph^{-1} defined on $\mathcal{B}(T)$ by $(Ph^{-1})(A) = P(h^{-1}(A))$ for all $A \in \mathcal{B}(T)$. Since h is Borel measurable, it follows immediately that Ph^{-1} is well defined. It is also easy to see that Ph^{-1} is indeed a probability measure on $(T, \mathcal{B}(T))$.

[6] If there were uncountably many such sets with positive measure, then infinitely many of them would have measure greater than $1/n$ for some $n \in \mathbb{N}$, so T would have infinite measure.

Theorem 1.1.6 (Continuous Mapping Theorem)**.** Under the above assumptions, let $P, \{P_n\}_{n\geq 1}$ be Borel probability measures on $\mathcal{B}(S)$. If $P_n \Rightarrow P$ and $P(D_h) = 0$, then $P_n h^{-1} \Rightarrow P h^{-1}$.

Proof. We will use the Portmanteau theorem. Let $F \subset T$ be a closed set. If $x \in \overline{h^{-1}(F)}$ then there exists a sequence $\{x_n\}$ in $h^{-1}(F)$ such that $x_n \to x$. From the definition of this sequence, it follows that $h(x_n) \in F$ for all indices n. If $x \in D_h^c$, then we have $h(x_n) \to h(x)$, so $h(x) \in \overline{F} = F$, which gives that $x \in h^{-1}(F)$. Thus, $D_h^c \cap \overline{h^{-1}(F)} \subset h^{-1}(F)$. Using the fact that $P(D_h^c) = 1$, we get

$$\begin{aligned}
\limsup_{n\to\infty} P_n(h^{-1}(F)) &\leq \limsup_{n\to\infty} P_n\left(\overline{h^{-1}(F)}\right) \\
&\leq P\left(\overline{h^{-1}(F)}\right) \\
&= P\left(D_h^c \cap \overline{h^{-1}(F)}\right) \\
&\leq P\left(h^{-1}(F)\right),
\end{aligned}$$

where the second inequality follows from the fact that $P_n \Rightarrow P$ and the Portmanteau theorem. The above series of inequalities show that condition (iii) of the Portmanteau theorem is satisfied, so $P_n h^{-1} \Rightarrow P h^{-1}$. □

1.1.2 Convergence in Distribution

The Portmanteau theorem shows that the notion of convergence we have defined indeed generalizes the notion of convergence in distribution. However, all the results we have proven so far are stated in terms of probability measures, while convergence in distribution is usually stated in terms of random variables. Paraphrasing these results using random variables instead of measures gives them a more familiar and concrete form.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(T, \mathcal{B}(T))$ a metric space with its Borel $\sigma-$field, and let $X : \Omega \to T$ be a random element, that is, a $\mathcal{A}/\mathcal{B}(T)-$measurable mapping. The *distribution* of X is the probability measure P_X on $\mathcal{B}(T)$ defined as $P_X := PX^{-1}$, that is,

$$P_X(A) = P(X^{-1}(A)) = P(X \in A).$$

The measurability of X ensures that P_X is well defined. It is easy to see that P_X is indeed a probability measure.

If $f : T \to \mathbb{R}$ is a Borel measurable function (that is, $\mathcal{B}(T)/\mathcal{B}(\mathbb{R})-$measurable), then by change of variable [Billingsley, 2008, p. 229], we get that

$$\mathbb{E}[f(X)] = \int_\Omega f(X(\omega))\, dP(\omega) = \int_T f(x)\, dP_X(x) = P_X f, \tag{1.5}$$

where the equality means that either both integrals exist or both do not exist, and if they exist, they have the same value.

Definition 1.1.7. Let $\{X_n\}$ be a sequence of random elements. We say that $\{X_n\}$ *converges in distribution* to a random element X if $P_{X_n} \Rightarrow P_X$. We shall then write that $X_n \Rightarrow X$.

Note that this definition only makes sense if $P_X, P_{X_1}, P_{X_2}, \ldots$ are defined on the same measurable space, that is, if the images of $X, X_1, X_2, \ldots$ are the same sets, with the same topology. However, we observe that the domains $(\Omega, \mathcal{A}, P), (\Omega_1, \mathcal{A}_1, P_1), \ldots$ of $X, X_1, \ldots$ respectively, that is, the underlying probability spaces, need not be equal.

We shall make the following convention: normally, we would write $\mathbb{E}_n$ for integrals with respect to the measure P_n, and $\mathbb{E}$ for integrals with respect to P. Equation (1.5) implies that $\mathbb{E}_n[f(X_n)] \to \mathbb{E}[f(X)]$ if and only if $P_{X_n} f \to P_X f$. From now on, instead of P_n and $\mathbb{E}_n$, we shall write P and $\mathbb{E}$, to refer to whatever underlying probability space the random element we are dealing with is defined on.

Using the definition of convergence in distribution and the above convention we get that $X_n \Rightarrow X$ if and only if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous functions $f : T \to \mathbb{R}$.

For convergence in distribution we have an analogue of the Portmanteau theorem. In what follows, an $X-$continuity set $A \in \mathcal{B}(T)$ is a set that satisfies the condition $P(X \in \partial A) = 0$. In other words, $P(X^{-1}(\partial A)) = 0$ or $P_X(\partial A) = 0$.

Theorem 1.1.8 (Portmanteau Theorem)**.** Let $X : (\Omega, \mathcal{A}, P) \to (T, \mathcal{B}(T))$ be a random element. The following conditions are equivalent:

(i) $X_n \Rightarrow X$,

(ii) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, uniformly continuous functions $f : T \to \mathbb{R}$,

(iii) $\limsup_n P(X_n \in F) \leq P(X \in F)$ for all closed sets $F \subset T$,

(iv) $\liminf_n P(X_n \in G) \geq P(X \in G)$ for all open sets $G \subset T$,

(v) $P(X_n \in A) \to P(X \in A)$ for all $X-$continuity sets $A \in \mathcal{B}(T)$.

This theorem is an immediate consequence of theorem 1.1.4 so we will omit its proof.

1.1.3 Tightness and Prohorov's Theorems

Prohorov's theorems are the second most important results in the theory of weak convergence, after the continuous mapping theorem. To state these theorems, we first need to introduce the notion of tightness.

Definition 1.1.9. Let $(T, \mathcal{B}(T))$ be a metric space together with its Borel $\sigma-$field. A probability measure P on $(T, \mathcal{B}(T))$ is called *tight* if, for any $\varepsilon > 0$, there exists a compact set $K \subset T$ such that $P(K) > 1 - \varepsilon$.

Theorem 1.1.10. If the metric space T is separable and complete, then any Borel probability measure on T is tight.

Proof. Let $k \geq 1$ be a positive integer, and $\varepsilon > 0$ an arbitrary positive real number. Since T is separable, there exists a countable set D that is dense in T. If we choose some arbitrary $t \in T$, then, since D is dense, there exists an element $d \in D$ whose distance from t is less than $1/k$. This shows that T can be covered by a countable family $A_1^{(k)}, A_2^{(k)}, \ldots$ of open balls of radius $1/k$ whose centers lie in D.

We consider the sequence of sets

$$B_n = \bigcup_{i=1}^{n} A_i^{(k)},\ n = 1, 2, \ldots.$$

This sequence is increasing, and

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} B_n\right) &= P\left(\bigcup_{i=1}^{\infty} A_i^{(k)}\right) \\ &= 1, \end{aligned}$$

so $\lim_{n\to\infty} P(B_n) = 1$. We choose $n_k \in \mathbb{N}$ such that $P(B_{n_k}) > 1 - \varepsilon/2^k$. We repeat this process for all positive integers k. Notice that the set

$$A = \bigcap_{k=1}^{\infty} \bigcup_{i=1}^{n_k} A_i^{(k)}$$

is totally bounded. Indeed, if $\delta > 0$ and k_δ is a positive integer with $1/k_\delta < \delta$, then the above set is covered by the (finitely many) open balls

$$A_1^{(k_\delta)}, A_2^{(k_\delta)}, \ldots, A_{n_{k_\delta}}^{(k_\delta)}.$$

We know that, if a set is totally bounded, then its closure is totally bounded too. Thus, $\overline{A}$ is totally bounded and, since T is complete, it follows that $\overline{A}$ is bounded. We have

$$\begin{aligned} P(A^c) &= P\left(\bigcup_{k=1}^{\infty}\left(\bigcup_{i=1}^{n_k} A_i^{(k)}\right)^c\right) \\ &\leq \sum_{k=1}^{\infty} P\left(\left(\bigcup_{i=1}^{n_k} A_i^{(k)}\right)^c\right) \\ &< \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} \\ &= \varepsilon, \end{aligned}$$

so $P(\overline{A}) \geq P(A) > 1 - \varepsilon$. □

Tightness of a family of probability measures is defined in a similar way:

Definition 1.1.11. Let T be a metric space and let $\mathcal{P}$ be a family of Borel probability measures on T. We say that $\mathcal{P}$ is *tight* if, for any $\varepsilon > 0$, there exists a compact set $K \subset T$ such that $P(K) > 1 - \varepsilon$ for all $P \in \mathcal{P}$.

Prohorov's theorems connect tightness with the notion of *relative compactness* defined as follows:

Definition 1.1.12. Let T be a metric space and let $\mathcal{P}$ be a family of Borel probability measures on T. We say that $\mathcal{P}$ is *relatively compact* if, for any sequence $\{P_n\}$ in $\mathcal{P}$, there exists a subsequence $\{P_{k_n}\}$ that converges weakly to a Borel probability measure Q.

Notice that the above definition requires Q to be defined on $(T, \mathcal{B}(T))$, but Q need not be an element of $\mathcal{P}$. This definition is analogous to the definition of sequential compactness of a subset of a metric space.

We are now ready to state Prohorov's theorems.

Theorem 1.1.13 (Prohorov)**.** If the family $\mathcal{P}$ is tight, then it is relatively compact.

Proof. The proof we present is due to Billingsley [Billingsley, 1971] and is split in seven parts. First, we consider a sequence $\{P_n\}$ in $\mathcal{P}$. We want to find a subsequence $\{P_{k_n}\}$ and a Borel probability measure P such that $P_{k_n} \Rightarrow P$.

PART 1: Due to the tightness of $\mathcal{P}$, we can find compact sets $\{K_u\}_{u \geq 1}$ such that $K_1 \subset K_2 \subset \ldots$ and $P_n(K_u) > 1 - u^{-1}$ for all u and n. The set $\bigcup_u K_u$ is separable, so there exists a countable class $\mathcal{A}$ of open sets with the property that, if x lies both in $\bigcup_u K_u$ and in $G \subset T$, and if G is open, then

$x \in A \subset \overline{A} \subset G$ for some $A \in \mathcal{A}$. Let $\mathcal{H}$ be the (countable) class containing $\emptyset$ and all the finite unions of sets of the form $\overline{A} \cap K_u$, where $A \in \mathcal{A}$ and $u \geq 1$. Note that each K_u belongs to $\mathcal{H}$. Indeed, from the definition of $\mathcal{A}$ it follows immediately that for each x that lies in K_u and in the open set $G = T$, there exists an open set $A_x \in \mathcal{A}$ such that $x \in A \subset$. Then, $\{A_x\}_{x \in K_u}$ is an open covering of K_u. Since K_u is compact, there exist $x_1, \ldots, x_n \in K_u$ such that

$$K_u \subset \bigcup_{j=1}^{n} A_{x_j}.$$

From the latter equation it follows that

$$K_u = \bigcup_{j=1}^{n} (\overline{A_{x_j}} \cap K_u),$$

so $K_u \in \mathcal{H}$.

Using the diagonal method and the fact that $\mathcal{H}$ is countable, we can find a subsequence $\{P_{n_i}\}$ for which the limit

$$a(H) := \lim_{i \to \infty} P_{n_i}(H) \tag{1.6}$$

exists for all $H \in \mathcal{H}$. Our purpose is to construct on $\mathcal{B}(T)$ a probability measure P such that

$$P(G) = \sup_{H \subset G} a(H) \tag{1.7}$$

for all open sets G. If such a measure exists, then the proof will be completed as follows: if G is an open set and $H \subset G$, then $P_{n_i}(H) \leq P_{n_i}(G)$ for $i = 1, 2, \ldots$ so

$$\begin{aligned} a(H) &= \lim_{i \to \infty} P_{n_i}(H) \\ &\leq \liminf_{i \to \infty} P_{n_i}(G). \end{aligned}$$

Using (1.7) we now get that $P(G) \leq \liminf_i P_{n_i}(G)$, so the Portmanteau theorem gives $P_{n_i} \Rightarrow P$. Thus, the existence of a measure satisfying (1.7) indeed proves the theorem.

To construct such a measure, we first observe that $\mathcal{H}$ is closed under finite unions. Also, it is easy to see that $a(H)$ has the following properties:

- $a(\emptyset) = 0$,
- $a(H_1) \leq a(H_2)$ if $H_1 \subset H_2$,

- $a(H_1 \cup H_2) = a(H_1) + a(H_2)$ if $H_1 \cap H_2 = \emptyset$,
- $a(H_1 \cup H_2) \leq a(H_1) + a(H_2)$.

For open sets G we define

$$\beta(G) = \sup_{H \subset G} a(H). \tag{1.8}$$

Then, β is monotone and $\beta(\emptyset) = a(\emptyset) = 0$. Finally, for any subset M of T we define

$$\gamma(M) = \inf_{M \subset G} \beta(G). \tag{1.9}$$

We observe that, if $G \subset T$ is an open set, then $\gamma(G) = \beta(G)$. Indeed,

$$\gamma(G) = \inf_{G \subset G'} \beta(G') \leq \beta(G),$$

and, for the converse inequality, we observe that, for all open sets $G' \supset G$ it is true that $\beta(G') \geq \beta(G)$, so

$$\gamma(G) = \inf_{G \subset G'} \beta(G') \geq \beta(G).$$

Our purpose is to show that γ is an outer measure. In this case, recall that a set $M \subset T$ will be $\gamma-$measurable if and only if $\gamma(L) \geq \gamma(M \cap L) + \gamma(M^c \cap L)$ (this follows immediately from the definition of Caratheodary for the measurability of a set). Also, from Caratheodary's theorem [Folland, 1999, p. 29] it follows that the class $\mathcal{M}$ of $\gamma-$measurable sets will be a $\sigma-$field and the restriction of γ to $\mathcal{M}$ will be a (complete) measure. Suppose that we are also able to show that each closed set lies in $\mathcal{M}$. Then, it will follow that $\mathcal{B}(T) \subset \mathcal{M}$ and that the restriction P of γ to $\mathcal{B}(T)$ is a measure satisfying $P(G) = \gamma(G) = \beta(G)$. Thus, equation (1.7) will hold. Note that P will be a probability measure as

$$1 \geq P(T) = \beta(T) \geq \sup_{u \geq 1} a(K_u) \geq \sup_{u \geq 1} (1 - u^{-1}) = 1.$$

For this last series of inequalities we used the fact that each K_u lies in $\mathcal{H}$, as we proved earlier. Thus, it suffices to show that γ is an outer measure and that each closed set is $\gamma-$measurable.

PART 2. In this part we will prove that if F is closed, G is an open set such that $F \subset G$, and if there exists $H \in \mathcal{H}$ such that $F \subset H$, then there exists $H_0 \in \mathcal{H}$ such that $F \subset H_0 \subset G$.

Note that $F \subset H$ and H is a union of sets of the form $\overline{A} \cap K_u$ $(A \in \mathcal{A})$, so F is certainly contained in $\bigcup_u K_u$. Thus, for each $x \in F \subset G$ we can choose a set $A_x \in \mathcal{A}$ such that $x \in A_x \subset \overline{A_x} \subset G$.

Also, we observe that each set of the form $\overline{A} \cap K_u \subset K_u$ is compact (closed subset of a compact set), so H is itself compact as a finite union of such sets. It follows that F is also compact because it is a closed subset of H. Our last observation is that, since $K_1 \subset K_2 \subset \ldots$, the set H must be contained in some K_u. The family $\{A_x\}_{x \in F}$ is an open cover of F, so it has a finite subcover $A_{x_1}, \ldots, A_{x_k}$. Since $F \subset H \subset K_u$ for some u, we can choose

$$H_0 = \bigcup_{i=1}^{k} \left(\overline{A_{x_i}} \cap K_u\right).$$

PART 3. In this part we will show that β is finitely subadditive (on the open sets). Let G_1, G_2 be open sets and suppose that $H \in \mathcal{H}$ satisfies $H \subset G_1 \cup G_2$. We consider the sets

$$F_1 = \{x \in H : \rho(x, G_1^c) \geq \rho(x, G - 2^c)\} \text{ and } F_2 = \{x \in H : \rho(x, G_2^c) \geq \rho(x, G - 1^c)\},$$

where ρ denotes the distance metric of T.

If $x \in F_1$ and $x \notin G_1$, then $x \in G_2$ because $F_1 \subset H \subset G_1 \cup G_2$. Since G_2^c is closed, we get that $\rho(x, G_2^c) > 0$ so $\rho(x, G_1^c) = 0 < \rho(x, G_2^c)$, contradiction. Thus, $F_1 \subset G_1$ and similarly $F_2 \subset G_2$. Note that F_1, F_2 are closed sets due to the continuity of the function ρ. Since $F_1 \subset H$, it follows from the previous part that $F_1 \subset H_1 \subset G_1$ for some $H_1 \in \mathcal{H}$. Similarly, $F_2 \subset H_2 \subset G_2$ for some $H_2 \in \mathcal{H}$. Then, since $H = F_1 \cup F_2$, the properties of a and equation (1.8) give

$$\begin{aligned} a(H) &= a(F_1 \cup F_2) \\ &\leq a(H_1 \cup H_2) \\ &\leq a(H_1) + a(H_2) \\ &\leq \beta(G_1) + \beta(G_2). \end{aligned}$$

Taking the supremum over all $H \subset G_1 \cup G_2$ yields that

$$\beta(G_1 \cup G_2) \leq \beta(G_1) + \beta(G_2).$$

PART 4. We will show that β is countably subadditive (on the open sets). Let $\{G_n\}_{n \geq 1}$ be a sequence of open sets, and let $H \in \mathcal{H}$ satisfy

$$H \subset \bigcup_{n=1}^{\infty} G_n.$$

Recall that H is compact (as an element of $\mathcal{H}$) so the open cover $\{G_n\}_{n\geq 1}$ of H will have a finite subcover. In particular, there exists $n_0 \geq 1$ such that

$$H \subset \bigcup_{n=1}^{n_0} G_n.$$

Finite subadditivity implies that

$$\begin{aligned} a(H) &\leq \beta\left(\bigcup_{n=1}^{n_0} G_n\right) \\ &\leq \sum_{n=1}^{n_0} \beta(G_n) \\ &\leq \sum_{n=1}^{\infty} \beta(G_n). \end{aligned}$$

Taking the supremum over all such sets H gives that

$$\beta\left(\bigcup_{n=1}^{\infty} G_n\right) \leq \sum_{n=1}^{\infty} \beta(G_n).$$

PART 5. We are now ready to show that γ is an outer measure. From its definition, it is easy to see that γ is monotone and that $\gamma(\emptyset) = 0$. It remains to show that it is countably subadditive. We consider an arbitrary $\varepsilon > 0$, and subsets $\{M_n\}_{n\geq 1}$ of T. By the definition of γ, we can find open sets $\{G_n\}_{n\geq 1}$ such that $\beta(G_n) < \gamma(M_n) + \varepsilon/2^n$ for all $n \geq 1$. Since β is countably subadditive, it follows that

$$\begin{aligned} \gamma\left(\bigcup_{n=1}^{\infty} M_n\right) &\leq \beta\left(\bigcup_{n=1}^{\infty} G_n\right) \\ &\leq \sum_{n=1}^{\infty} \beta(G_n) \\ &< \varepsilon + \sum_{n=1}^{\infty} \gamma(M_n). \end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, it follows that

$$\gamma\left(\bigcup_{n=1}^{\infty} M_n\right) \leq \sum_{n=1}^{\infty} \gamma(M_n),$$

so γ is indeed an outer measure.

PART 6. It remains to show that all closed sets are $\gamma-$measurable. We are first going to show that

$\beta(G) \geq \gamma(F \cap G) + \gamma(F^c \cap G)$ for all closed sets F and open sets G. We first consider $\varepsilon > 0$. The set $F^c \cap G$ is open so, from the definition of β, we can choose $H_1 \in \mathcal{H}$ such that $H_1 \subset F^c \cap G$ and $a(H_1) > \beta(F^c \cap G) - \varepsilon$. Also, the set H_1 is compact, so $H_1^c \cap G$ is open. Thus, just like before, we can find $H_0 \in \mathcal{H}$ such that $H_0 \subset H_1^c \cap G$ and $a(H_0) > \beta(H_1^c \cap G) - \varepsilon$. Since H_0 and H_1 are disjoint and are contained in G, it follows from the definitions of β, γ and from the properties of α that

$$\begin{aligned}
\beta(G) &\geq a(H_0 \cup H_1) \\
&= a(H_0) + a(H_1) \\
&> \beta(H_1^c \cap G) + \beta(F^c \cap G) - 2\varepsilon \\
&\geq \gamma(F \cap G) + \gamma(F^c \cap G) - 2\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the desired inequality follows.

PART 7. We consider a closed set $F \subset T$ and set $L \subset T$. From the previous part we get that $\beta(G) \geq \gamma(F \cap L) + \gamma(F^c \cap L)$ for all open sets G such that $L \subset G$. Taking the infimum over all those G gives us that

$$\gamma(L) \geq \gamma(F \cap L) + \gamma(F^c \cap L),$$

so F is indeed $\gamma-$measurable, which completes the proof of the theorem. □

The above theorem has the following useful corollary:

Corollary 1.1.14. Let T be a metric space and let P be a Borel probability measure on T. Suppose that the sequence $\{P_n\}$ of Borel probability measures is tight, and that the limit of any weakly convergent subsequence is P. Then, $P_n \Rightarrow P$.

Proof. Since $\{P_n\}$ is tight, it follows from Prohorov's theorem that it is also relatively compact. Thus, each subsequence has a further subsequence that is weakly convergent. From our hypothesis we get that the limit of this further subsequence is P. Using theorem 1.1.5, we get that $P_n \Rightarrow P$. □

Under some additional assumptions, the converse implication of Prohorov's theorem is also true. This is shown in the next theorem.

Theorem 1.1.15 (Prohorov)**.** Suppose that T is separable and complete. If $\mathcal{P}$ is relatively compact, then it is tight.

Proof. We consider an increasing sequence $\{G_n\}_{n\geq 1}$ with $\bigcup_n G_n = T$. We will first show the following assertion: for all $\varepsilon > 0$, there exists an integer $n \geq 1$ such that $P(G_n) > 1 - \varepsilon$ for all $P \in \mathcal{P}$. If this

were not true, then for each n there would exist $P_n \in \mathcal{P}$ such that $P_n(G_n) \leq 1 - \varepsilon$. From the relative compactness of the family $\{P_n\}$, there exists a subsequence $\{P_{n_i}\}_{i\geq 1}$ and a Borel probability measure Q on T such that $P_{n_i} \Rightarrow Q$. From the Portmanteau theorem it follows that

$$Q(G_n) \leq \liminf_{i\to\infty} P_{n_i}(G_n) \leq \liminf_{i\to\infty} P_{n_i}(G_{n_i}) \leq 1 - \varepsilon$$

for all $n \geq 1$, where the second inequality follows from the fact that $\{G_n\}_{n\geq 1}$ is increasing and $n_i \to \infty$. However, the above relation cannot be true since $G_n \nearrow T$, so $Q(G_n) \to Q(T) = 1$.

From the above assertion it follows that, if $k \geq 1$ and if $A_1^{(k)}, A_2^{(k)}, \ldots$ is a sequence of open balls of radius $1/k$ covering T (such balls exists due to the fact that T is separable - we have used this argument in theorem 1.1.10), then there exists $n_k \geq 1$ such that

$$P\left(\bigcup_{i=1}^{n_k} A_i^{(k)}\right) > 1 - \frac{\varepsilon}{2^k} \text{ for all } P \in \mathcal{P}.$$

Indeed, the sequence $\{G_n^{(k)}\}_{n\geq 1}$ defined by

$$G_n = \bigcup_{i=1}^{n} A_i^{(k)}$$

is increasing and $\bigcup_{n=1}^{\infty} G_n = T$, so the above assertion can be applied. Just like in the proof of the theorem 1.1.10, the set

$$A = \bigcap_{k=1}^{\infty} \bigcup_{i=1}^{n_k} A_{ki}$$

is totally bounded, it has compact closure, and $P\left(\overline{A}\right) > 1 - \varepsilon$ for all $P \in \mathcal{P}$. Thus, $\mathcal{P}$ is tight. □

1.1.4 Examples

Before concluding the section, we are going to look at two important examples that will help us get some intuition about the nature of the spaces $\mathbb{R}^\infty$ (sequences of real numbers) and $C = C[0,1]$ (space of continuous functions $[0,1] \to \mathbb{R}$). Although these examples are a bit technical, they will prove very useful for our analysis of weak convergence on $C[0,1]$.

We say that a subclass $\mathcal{A}$ of the Borel $\sigma-$field $\mathcal{B}(T)$ of a metric space T is a *separating class* if, for any Borel probability measure P on T, the values $P(A),\ A \in \mathcal{A}$ completely determine P. In other words, $\mathcal{A}$ is a separating class if, any two Borel probability measures P, Q on T that agree on each $A \in \mathcal{A}$ are equal.

Example 1.1.16. Let $\mathbb{R}^\infty$ be the space of sequences $x = (x_1, x_2, \ldots)$ of real numbers. We consiδer the metric b on $\mathbb{R}$ defined by $b(\alpha, \beta) = \min(1, |\alpha - \beta|)$. This metric is equivalent to the usual one and, with this metric, $\mathbb{R}$ is a complete separable metric space. We consider the metric ρ on $\mathbb{R}^\infty$ defined by

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

Notice that ρ is well defined because b is bounded. Also, it is easy to check that ρ is indeed a metric. Obviously, if $\rho(x^n, x) \to_n 0$, then $b(x_i^n, x_i) \to_n 0$ for all i. We shall show that the converse is also true. Suppose that $b(x_i^n, x_i) \to_n 0$ for $i = 1, 2, \ldots$ and take $\varepsilon > 0$. There exists $N \geq 1$ such that $\sum_{i \geq N} 1/2^n < \varepsilon/2$. Since $b(\alpha, \beta) \leq 1$ for all $\alpha, \beta \in \mathbb{R}$, it follows that $\sum_{i \geq N} b(x_i^n, x_i)/2^i < \varepsilon/2$ for all $n \geq 1$. Also, from the condition $b(x_i^n, x_i) \to_n 0$, it follows that there exists an integer $n_\varepsilon \geq 1$ such that $b(x_i^n, x_i) < \varepsilon/2$ for all $i < N$ and for all $n \geq n_\varepsilon$ (we have finitely many i, so it is indeed possible to find a unique n_ε that works for all of them). Thus, for all $n \geq n_\varepsilon$ we have

$$\begin{aligned} \rho(x^n, x) &= \sum_{1 \leq i < N} \frac{b(x_i^n, x_i)}{2^i} + \sum_{i \geq N} \frac{b(x_i^n, x_i)}{2^i} \\ &\leq \frac{\varepsilon}{2} \sum_{1 \leq i < N} \frac{1}{2^i} + \frac{\varepsilon}{2} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

which gives that $\rho(x^n, x) \to_n 0$. Thus, with the metric ρ, the space $\mathbb{R}^\infty$ has the topology of pointwise convergence: $x^n \to x$ if and only if $x_i^n \to x_i$ for $i = 1, 2, \ldots$

Let $\pi_k : \mathbb{R}^\infty \to \mathbb{R}^k$ be the natural projection: $\pi_k(x) = (x_1, \ldots, x_k)$. If we assume that $\mathbb{R}^k$ has its usual Euclidean metric, then it is easy to see that π_k is continuous. Indeed, if $x^n \to x$, then $x_i^n \to x_i$ for all i, so

$$(x_1^n, \ldots, x_k^n) \to (x_1, \ldots, x_k),$$

or, in other words, $\pi_k(x^n) \to \pi_k(x)$. From the continuity of π_k it follows that the sets

$$N_{k,\varepsilon}(x) = \{y : \ |y_i - x_i| < \varepsilon, \ i = 1, \ldots, k\}$$

are open, as they are the inverse images of open sets under π_k:

$$N_{k,\varepsilon}(x) = \pi_k^{-1} \left(\prod_{i=1}^{k} (x_i - \varepsilon, x_i + \varepsilon) \right). \tag{1.10}$$

Moreover, if $y \in N_{k,\varepsilon}(x)$, then $\rho(x, y) < \varepsilon + 2^{-k}$. Given a positive radius r, we can choose k and

ε such that $\varepsilon + 2^{-k} < r$. Then, it is true that $N_{k,\varepsilon}(x) \subset B(x,r)$, so the sets $N_{k,\varepsilon}(x)$ form a base for the topology of $\mathbb{R}^\infty$. We observe that the space is separable: a countable, dense subset consists of those points having only finitely many nonzero coordinates, each of them rational. Also, if $\{x^n\}$ is a Cauchy sequence, then it is easy to see that each $\{x_i^n\}$ is Cauchy too. Thus, $x_i^n \to_n x_i$, so $\{x^n\}$ converges to the sequence $x = (x_1, x_2, \ldots)$. It follows that $\mathbb{R}^\infty$ is also complete.

Let $\mathcal{R}_f^\infty$ be the class of *finite-dimensional* set, that is, the sets of the form $\pi_k^{-1}(H)$ for some $k \geq 1$ and some $H \in \mathcal{B}(\mathbb{R}^k)$. Since π_k is continuous, it is $\mathcal{B}(\mathbb{R}^\infty)/\mathcal{B}(\mathbb{R}^k)-$measurable, so $\mathcal{R}_f^\infty \subset \mathcal{B}(\mathbb{R}^\infty)$. Note that $\pi_k^{-1}(H) = \pi_{k+1}^{-1}(H \times \mathbb{R})$ so any two sets $A, A' \in \mathcal{R}_f^\infty$ can be written as $\pi_k^{-1}(H)$ and $\pi_k^{-1}(H')$ respectively, for the same value of k. It follows that $A \cap A' = \pi_k^{-1}(H \cap H')$ so $\mathcal{R}_f^\infty$ is a $\pi-$system.[7] Since the sets $N_{k,\varepsilon}(x)$ form a base and each of them lies in $\mathcal{R}_f^\infty$ (see equation (1.10)), it follows by separability that each open set is a countable union of sets in $\mathcal{R}_f^\infty$. Thus, $\mathcal{R}_f^\infty$ generates the Borel $\sigma-$field $\mathcal{B}(\mathbb{R}^\infty)$. Since $\mathcal{R}_f^\infty$ is also a $\pi-$system, it follows that it is a separating class.

If P is a Borel probability measure on $\mathbb{R}^\infty$, its *finite dimensional distributions* are the Borel probability measures $P\pi_k^{-1}$[8] on $\mathbb{R}^k$, $k \geq 1$. Since $\mathcal{R}_f^\infty$ is a separating class, these measures completely determine P.

Example 1.1.17. We will now deal with the space $C = C[0,1]$. We denote the uniform metric on this space by ρ. In other words, $\rho(x,y) = \|x - y\|_\infty$. Note that, if $\rho(x_n, x) \to 0$, then x_n converges uniformly to x, so it also converges pointwise.[9]

The space C is separable. Let D_k be the set of *polygonal functions* that are linear over each subinterval $I_{ki} = [(i-1)/k, i/k]$ and have rational values at the endpoints. Then, each D_k is countable, so $\bigcup_k D_k$ is countable. To show that it is dense, we consider $x \in C[0,1]$ and $\varepsilon > 0$. Notice that x is uniformly continuous, so it is possible to find $k \geq 1$ such that $|x(t) - x(i/k)| < \varepsilon$ for all $t \in I_{ki}$, $1 \leq i \leq k$. We can now choose $y \in D_k$ such that $|y(i/k) - x(i/k)| < \varepsilon$ for $1 \leq i \leq k$. We have that $|y(i/k) - x(t)| \leq |y(i/k) - x(i/k)| + |x(i/k) - x(t)| \leq 2\varepsilon$ for all $t \in I_{ki}$, and that $|y((i-1)/k) - x(t)| \leq |y((i-1)/k) - x((i-1)/k)| + |x((i-1)/k) - x(i/k)| + |x(i/k) - x(t)| \leq 3\varepsilon$ for all $x \in I_{ki}$. From the construction of y, it follows that $y(t)$ is a convex combination of $y((i-1)/k)$ and $y(i/k)$ for $t \in I_{ki}$.[10] Thus, the above inequalities imply that $|y(t) - x(t)| < 3\varepsilon$ for $t \in I_{ki}$.[11] Thus, $\rho(x,y) \leq 3\varepsilon$, which shows that $\bigcup_k D_k$ is dense.

[7] A $\pi-$system is a non empty family $\mathcal{L}$ of subsets of a set X that is closed under finite intersections.

[8] You can look at the paragraph before theorem 1.1.6 to remember how $P\pi_k^{-1}$ is defined.

[9] The converse is of course not true.

[10] This happens because $y(t)$ lies on the segment with endpoints $y((i-1)/k)$ and $y(i.k)$.

[11] If $y(t) = ay((i-1)/k) + by(i/k)$ with $a, b \geq 0$ and $a + b = 1$, then $|y(t) - x(t)| \leq a|y((i-1)/k) - x(t)| + b|y(i/k) - x(t)| \leq 2a\varepsilon + 3b\varepsilon \leq 3\varepsilon(a+b) = 3\varepsilon$.

We are going to show that C is complete too. If $\{x_n\}$ is a Cauchy sequence in C, then, for each $t \in [0,1]$, the sequence $\{x_n(t)\}$ of real numbers is Cauchy too, so it has a limit $x(t)$. Since $\{x_n\}$ is Cauchy, it follows that $\varepsilon_n = \sup_{m>n} \rho(x_m, x_n) \to_n 0$. We have $|x_n(t) - x_m(t)| \leq \varepsilon_n$ and, letting $m \to \infty$ gives $|x_n(t) - x(t)| \leq \varepsilon_n$. We take the supremum over all $t \in [0,1]$ and we get that

$$\rho(x_n, x) \leq \varepsilon_n,$$

whence it follows that x_n converges uniformly to x. Thus, x is continuous and $\rho(x_n, x) \to_n 0$, so C is complete.

For $0 \leq t_1 < t_2 < \ldots < t_k \leq 1$, define the natural projection from C to $\mathbb{R}^k$ by $\pi_{t_1,\ldots,t_k}(x) = (x(t_1), \ldots, x(t_k))$. In C, the finite dimensional sets are those of the form $\pi^{-1}_{t_1,\ldots,t_k}(H)$ for $H \in \mathcal{B}(\mathbb{R}^k)$, and they lie in the Borel $\sigma-$field $\mathcal{B}(C)$ because $\pi_{t_1,\ldots,t_k}$ are obviously continuous. Just like in the previous example, the index set defining a finite dimensional set can always be enlarged. For example, suppose that we want to enlarge t_1, t_2 to t_1, s, t_2, where $t_1 < s < t_2$. For the projection $\psi : \mathbb{R}^3 \to \mathbb{R}^2$ defined by $\psi(u, v, w) = (u, w)$ we have $\pi_{t_1,t_2} = \psi \circ \pi_{t_1,s,t_2}$ so $\pi^{-1}_{t_1,t_2}(H) = \pi^{-1}_{t_1,s,t_2}\left(\psi^{-1}(H)\right)$, and of course $\psi^{-1}(H) \in \mathcal{B}(\mathbb{R}^k)$ for all $H \in \mathcal{B}(\mathbb{R}^2)$ because ψ is continuous. The proof for the general cases follows the exact same idea. Like in the previous example, the enlargement property of the index sets allows us to show that the class $\mathcal{C}_f$ of finite dimensional sets of C is a $\pi-$system. Also, it is easy to see that

$$\overline{B(x,\varepsilon)} = \bigcap_{r \in \mathbb{Q} \cap [0,1]} \{y : \ |y(r) - x(r)| \leq \varepsilon\}.$$

Thus, the $\sigma-$field generated by $\mathcal{C}_f$ contains the closed balls, so, due to separability, it contains all the open sets. Since $\mathcal{C}_f$ is a $\pi-$system, it follows that $\mathcal{C}_f$ is a separating class.

1.1.5 Measurability Constraints

In the previous sections, we presented some fundamental theorems of the classical theory of weak convergence. Throughout our analysis, we assumed that X_n were Borel measurable mappings from some probability space $(\Omega_n, \mathcal{A}_n, P_n)$ to a metric space $(T, \mathcal{B}(T))$. This essentially requires that $X_n^{-1}(A) \in \mathcal{A}$ for all $A \in \mathcal{B}(T)$. In most cases, this measurability requirement is satisfied when T is a separable space. However, as we are going to see, it fails even for very simple choices of the map X_n when T is nonseparable. This problem was initially identified in 1965 by Chibsov [Chibisov, 1965] and was well analyzed in 1968 by Billingsley [Billingsley, 1968]. Billingsley's example was essentially the following one:

Let $\xi_1, \xi_2, \ldots \sim \text{Unif}(0,1)$ be independent random variables defined on some probability space $(\Omega, \mathcal{A}, P)$. We have defined the empirical distribution function of $\xi_1, \xi_2, \ldots$ as the collection of random variables $\{\mathbb{F}_n(t,\cdot)\}_{t\in[0,1]}$, where

$$\mathbb{F}_n(t,\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\xi_i(\omega)\leq t\}}.$$

The *uniform empirical process* $\{\mathbb{G}_n(t,\cdot)\}_{t\in[0,1]}$ is defined as

$$\mathbb{G}_n(t,\cdot) = \sqrt{n}(\mathbb{F}_n(t,\cdot) - t).$$

omega Both the empirical distribution function and the uniform empirical process are of primal importance for the theory of empirical processes. The theorems of Glivenko-Cantelli and Donsker give some insight on the behaviour of these two quantities, so the latter will appear very often in our analysis. It is therefore reasonable to require that the empirical distribution function and the uniform empirical process have "good" properties.

We observe that, for each $\omega \in \Omega$, the quantities $\mathbb{F}_n(t,\omega)$ and $\mathbb{G}_n(t,\omega)$ are functions of t, so we would like to have that they are random elements of a function space, so that we can talk about weak convergence of these quantities.

The first space that comes to our mind is $C = C[0,1]$, equipped with its uniform metric. If the empirical distribution function and the uniform empirical process were elements of this space, then they would automatically be $\mathcal{A}/\mathcal{B}(C)-$measurable. Indeed, as we have shown in the previous section, an element $X : (\Omega, \mathcal{A}, P) \to C[0,1]$ is Borel measurable if and only if the projections $X_t = \pi_t \circ X : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ are all Borel measurable. This requirement is fulfilled here, as $\mathbb{F}_n(t,\cdot)$ is Borel measurable for all t due to the measurability of $\xi_1, \xi_2, \ldots$ Thus, the empirical distribution function would be measurable, whence it would follow that the uniform empirical process would also be measurable.

Unfortunately, the above scenario is not true since $\mathbb{F}_n(t,\cdot), \mathbb{G}_n(t,\cdot)$ are not continuous functions of t. We can see this directly from the defintion of $\mathbb{F}_n(t,\omega)$. However, it is immediate that $\mathbb{F}_n(t,\cdot)$ is right-continuous and that it has a left limit everywhere. The functions with these two properties are called *càdlàg functions*, standing for "continue à gauche, limite à droite", and the space containing all càdlàg functions $f : [0,1] \to \mathbb{R}$ is denoted by $\mathbb{D}[0,1]$ or simply by $\mathbb{D}$.

It follows that each element of $\mathbb{D}$ is a bounded function, so $\mathbb{D}$ can be equipped with the uniform metric, just like $C[0,1]$. The essential difference of these two metric spaces is that $\mathbb{D}$ is not separable. Indeed, $\{\mathbf{1}_{[0,t)}\}_{t\in(0,1]}$ is an uncountable set of isolated points of $\mathbb{D}$.

The next example illustrates the nonmeasurability problem discussed above. More specifically, it

shows that, if $\mathbb{D}$ is equipped with the uniform metric and with the Borel $\sigma-$field induced by this metric, then the empirical distribution function (as well as the uniform empirical process) is not $\mathcal{A}/\mathcal{B}(\mathbb{D}[0,1])-$measurable.

Example 1.1.18. For convenience, let us assume that $n = 1$. Then, $\mathbb{F}_1(t,\omega) = \mathbf{1}_{\{\xi_1(\omega)\leq t\}}$. We shall show that $\mathbb{F}_1 : \Omega \to \mathbb{D}$ [12] is not $\mathcal{A}/\mathcal{B}(\mathbb{D})-$measurable. Notice that, for any $s \in [0,1]$, the function $\mathbf{1}_{[s,1]}(t)$, $t \in [0,1]$ belongs to $\mathbb{D}$. We consider the open ball $B_s := B(\mathbf{1}_{[s,1]}, 1/2)$. For each subset A of $[0,1]$, we consider the set

$$N_A = \bigcup_{s\in A} B_s.$$

This set is open as a union of open sets. Notice that $\{\omega \in \Omega : \mathbb{F}_1(\cdot,\omega) \in N_A\} = \{\omega \in \Omega : \xi_1(\omega) \in A\}$. Indeed, if $\xi_1(\omega) \in A$, then $\mathbb{F}_1(\cdot,\omega) = \mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) = \mathbf{1}_{[\xi_1(\omega),1]}(\cdot) \in B_{\xi_1(\omega)} \subset N_A$, which shows the first inclusion. For the converse inclusion, assume that $\mathbb{F}_1(\cdot,\omega) \in N_A$. Then, $\mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) \in N_A$ so $\mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) \in B_s$ for some $s \in A$. This means that $\left\|\mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) - \mathbf{1}_{[s,1]}\right\|_\infty < 1/2$, which implies that $\mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) = \mathbf{1}_{[s,1]}$ [13]. Since $\mathbf{1}_{[0,\cdot]}(\xi_1(\omega)) = \mathbf{1}_{[\xi_1(\omega),1]}(\cdot)$, it follows that $\xi_1(\omega) = s \in A$. Suppose that $\mathbb{F}_1$ is $\mathcal{A}/\mathcal{B}(\mathbb{D})-$measurable. Then $\{\omega \in \Omega : \mathbb{F}_1(\cdot,\omega) \in N_A\} \in \mathcal{A}$, so $\{\omega \in \Omega : \xi_1(\omega) \in A\} \in \mathcal{A}$. Thus, it would be possible to define a probability measure μ on the power set of $[0,1]$ by

$$\mu(A) = P\left(\{\xi_1 \in A\}\right) = P\left(\{\omega \in \Omega : \xi_1(\omega) \in A\}\right).$$

Notice that, for any interval I we have $\mu(I) = P(\{\xi_1 \in I\}) = \ell(I)$, where $\ell(I)$ denotes the length of I. The last equality in the above relation follows from the fact that ξ_1 is uniformly distributed. Hence, μ coincides with the Lebesgue meausre on the subintervals of $[0,1]$. However, the Borel $\sigma-$field of $[0,1]$ is generated by the family of (open/closed) intervals, so μ is identically equal to the Lebesgue measure on the Borel $\sigma-$field of $[0,1]$. Since μ is defined on the whole power set of $[0,1]$, it follows that it is an extension of the Lebesgue measure on this set. However, it is well known that such an extension does not exist. Hence, $\mathbb{F}_1$ is not $\mathcal{A}/\mathcal{B}(\mathbb{D})-$measurable.

Several attempts have been made to deal with this gap and develop a more complete weak convergence theory. In particular, [Skorokhod, 1956] and [Billingsley, 1968] came up with a separable topology for $\mathbb{D}[0,1]$, different from the one induced by the uniform metric. Billingsley also showed that this topology is induced by a metric, under which $\mathbb{D}[0,1]$ is complete. Due to the separability of $\mathbb{D}[0,1]$, arguments like those used for $C[0,1]$ can be applied, and the results from the classical theory of weak convergence are valid. For instance, under the new topology and the Borel $\sigma-$field generated by the corresponding metric, an element $X : (\Omega, \mathcal{A}, P) \to \mathbb{D}[0,1]$ is Borel measurable if and only

[12] Recall that $(\Omega, \mathcal{A}, P)$ is the space in which ξ_i are defined.

[13] If they took different values in at least one point, then their distance would be at least 1.

if the projections $\pi_t \circ X : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ are all Borel measurable. Recall that this result is true for $C[0,1]$, but not for $\mathbb{D}[0,1]$ with the uniform topology, which created the nonmeasurability problem discussed above.

Another approach was developed by [Dudley, 1966],[Dudley, 1967]. Dudley worked on $\mathbb{D}[0,1]$ with the standard (uniform) metric, but tried to apply the results of the classical theory of weak convergence using a smaller $\sigma-$field instead of the Borel $\sigma-$field. His idea was that the Borel $\sigma-$field is too large, so the measurability requirements, which usually translate as $X^{-1}(A) \in \mathcal{A}$ *for all Borel sets* A, fail [14]. With a smaller $\sigma-$field, the above relation would be required to hold for fewer sets A, so it would be easier to satisfy. Dudley used the *ball* $\sigma-$*field*, the $\sigma-$field generated by the *open balls* instead of all the open sets. In nonseparable spaces, this field is strictly contained in the Borel $\sigma-$field. Dudley's approach deals effectively with the uniform empirical process and with some simple extensions, but it fails for more general versions of the uniform empirical process, like the general empirical process which will be defined later.

Pyke and Shorack proposed a different method in 1968 [Pyke and Shorack, 1968]. While in the classical theory we say that $X_n \Rightarrow X$ when

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \tag{1.11}$$

for all bounded, continuous functions $f : T \to \mathbb{R}$ [15], Pyke and Shorack proposed that we require the latter relation only for functions f such that $f(X_n)$ is a measurable map from $(\Omega_n, \mathcal{A}_n)$ to $\mathbb{R}$ [16].

However, the most effective method was that of [Hoffmann-Jørgensen, 1991, Chapter 7]), which does not make use of any measurability assumptions. For this reason, it is necessary to replace measures with *outer measures*, and use (1.11) with *outer expectations* instead of expectations. This approach is presented in detail in [Van Der Vaart and Wellner, 1996].

1.2 Weak Convergence on Spaces of Continuous Functions

In the previous section we stated all our results for the general case of a metric space T. However, it may be clear from chapter 1 that are primarily interested in stochastic processes. The key observation is that stochastic processes can sometimes be described as random elements of a function space.

[14] Here $X : (\Omega, \mathcal{A}, P) \to \mathbb{D}[0,1]$.
[15] Here, T is the metric space in which X_n and X take values.
[16] $(\Omega_n, \mathcal{A}_n)$ is the measurable space in which X_n is defined.

Indeed, if $\{X_t\}_{t\in T}$ is a stochastic process indexed by a metric space T and Ω is its underlying probability space, then, for each $\omega \in \Omega$, $X_t(\Omega)$ is a measurable function from T to $\mathbb{R}$. In other words, $X_t(\omega) \in \mathbb{R}^T$.

Therefore, if we equip $\mathbb{R}^T$ with a distance metric such that the mapping $X : \Omega \to \mathbb{R}^T$ is Borel measurable, then X can be considered a random element of $\mathbb{R}^T$ and all the results of the previous section can be applied.

However, the space $\mathbb{R}^T$ contains too many functions, so it would be very difficult to extract additional results for this space. Furthermore, for almost all stochastic processes $\{X_t\}_{t\in T}$ appearing in our applications, the function $X_t(\omega) : T \to \mathbb{R}$ has extra properties (i.e. it may be bounded or continuous). Thus, it is unnecessary to work in a space like $\mathbb{R}^T$, because, apart from the fact that it is difficult to work with, it contains too many functions that are not of so much interest. On the other hand, working with smaller spaces like $\ell^\infty(T)$, the space of bounded functions $T \to \mathbb{R}$, or $C(T)$, the space of continuous functions $T \to \mathbb{R}$ is more fruitful.[17] In this section, we will work with the second space, $C(T)$, and we will choose $T = [0,1]$. We will denote this space by $C[0,1]$ or C.

This space has a very natural topology, induced by the distance function $\|\cdot\|_\infty$, which is defined as follows: if $x, y \in C[0,1]$, then

$$\|x - y\|_\infty = \sup_{t\in[0,1]} |x(t) - y(t)| .$$

Notice that x, y are bounded, so $\|\cdot\|$ is well defined. It is easy to verify that $\|\cdot\|$ is indeed a distance metric.

The main purpose of this section is to prove Donsker's Uniform Central Limit theorem, which is a refinement of the Lindeberg-Lévy Central Limit Theorem. The latter implies that, if $\xi_1, \xi_2, \ldots$ is a sequence of i.i.d. random variables defined on some probability space $(\Omega, \mathcal{A}, P)$, such that $\mathbb{E}(\xi_n) = 0$ and $\mathrm{Var}(\xi_n) = \sigma^2$, then

$$\frac{1}{\sigma\sqrt{n}} S_n = \frac{1}{\sigma\sqrt{n}} (\xi_1 + \ldots + \xi_n)$$

converges weakly to the normal distribution $\mathcal{N}(0,1)$.

For Donsker's theorem, we consider a point $\omega \in \Omega$ and we construct on $[0,1]$ the polygonal function that is linear in each of the intervals $[(i-1)/n, i/n]$ and takes the value $S_i(\omega)/(\sigma\sqrt{n})$ at the point i/n ($S_0(\omega) = 0$). In other words, we construct the function $X^n(\omega) : [0,1] \to \mathbb{R}$ whose value at $t \in [0,1]$ is

$$X_t^n(\omega) = \frac{1}{\sigma\sqrt{n}} S_{i-1}(\omega) + \frac{t - (i-1)/n}{1/n} \cdot \frac{1}{\sigma\sqrt{n}} \xi_i(\omega), \text{ for } t \in \left[\frac{i-1}{n}, \frac{i}{n}\right].$$

[17] A stochastic process $\{X_t\}_{t\in T}$ for which $X_t(\omega) : T \to \mathbb{R}$ is continuous for all $\omega \in \Omega$ is said to have *continuous sample paths*. Similarly, if $X_t(\omega) : T \to \mathbb{R}$ is bounded for all $\omega \in \Omega$, we say that the process has *bounded sample paths*.

From the above construction it follows that $X^n(\omega)$ is a continuous function on $[0,1]$. We will prove that $X^n : \Omega \to C$ is $\mathcal{A}/\mathcal{B}(C)-$measurable, so we can consider its distribution, that is, a Borel probability measure P_n on C, defined by

$$P_n(A) = P\left(\{\omega \in \Omega : X^n(\omega) \in A\}\right)$$

for all $A \in \mathcal{B}(C)$. Donsker's theorem states that

$$P_n \Rightarrow W,$$

where W is the *Wiener measure*, defined as the distribution of the Wiener process.[18] Showing the existence of the Wiener measure is also one of the purposes of this section.

However, before moving on to the proofs of these results, we need some technical tools regarding the structure of the compact sets in C, as well as some results about tightness in this space.

1.2.1 Weak Convergence and Tightness in C[0,1]

In example 1.1.17, we saw that the class $\mathcal{C}_f$ of finite dimensional sets of C is a separating class. Thus, by definition, each Borel probability measure on C is completely characterized by its values on the finite dimensional sets. One question that arises at this point is whether we can use the finite dimensional distributions of a measure (see example 1.1.16 for the definition) to get results about weak convergence. In particular, let $P, \{P_n\}_{n\geq 1}$ be Borel probability measures on C.[19] Assume that, for all index sets $\{0 \leq t_1 < \ldots < t_k \leq 1\}$ it is true that

$$P_n \pi^{-1}_{t_1,\ldots,t_k} \Rightarrow P\pi^{-1}_{t_1,\ldots,t_k}.$$

Does this yield that $P_n \Rightarrow P$? It would be very natural to assume that, since the finite dimensional distributions of a measure characterize it completely, the above conclusion is true. However, as the following example shows, this is not always the case.

Example 1.2.1. We consider the sequence $\{z_n\}_{n\geq 1}$ in C, which is defined as follows: z_n increases linearly from 0 to 1 over $[0, 1/n]$, decreases linearly from 1 to 0 over $[1/n, 2/n]$ and stays at 0 after $2/n$. In other words,

$$z_n(t) = nt\mathbf{1}_{[0,1/n]}(t) + (2 - nt)\mathbf{1}_{(1/n,2/n]}(t).$$

[18] The Wiener process is a random element of $C[0,1]$ whose properties were given in section 1

[19] Before moving on, the reader should stop at this point and make sure they understand what the last phrase means. It is important to remember that a probability measure on C is a measure that is defined on *sets of functions* and more specifically, on the Borel sets induced by the uniform metric.

Also, let $z : [0,1] \to \mathbb{R}$ be the zero function. We consider the restrictions of Dirac probability measures $P_n = \delta_{z_n}$ and $P = \delta_z$ to the Borel $\sigma-$field $\mathcal{B}(C)$.[20] It is easy to show that $P_n \Rightarrow P$ if and only if $z_n \overset{\|\cdot\|_\infty}{\to} z = 0$. However, $\|z_n - z\|_\infty = 1$ for all $n \geq 1$, so the above relation is not true, which gives that $P_n \not\Rightarrow P$.

On the other hand, if $2/n$ is less than the smallest non zero t_i, then $\pi_{t_1,\ldots,t_k}(z_n) = \pi_{t_1,\ldots,t_k}(0,\ldots,0)$, so $P_n\pi^{-1}_{t_1,\ldots,t_k}(H) = P\pi^{-1}_{t_1,\ldots,t_k}(H)$ for n large enough. Thus, in this case, the finite dimensional distributions of P_n converge weakly to those of P, while $P_n \not\Rightarrow P$. This counterexample indicates that, in the space C the arguments and results involving weak convergence go far beyond the finite dimensional theory.

Example 1.2.2. In the previous example we saw that weak convergence of the finite dimensional distributions is not strong enough to give weak convergence of the measures themselves. In this example, we will strengthen our hypothesis so that the above conclusion become true.

More specifically, we consider Borel probability measures $P, \{P_n\}_{n\geq 1}$ on C, and we assume that the finite dimensional distributions of P_n converge weakly to those of P. Here, we will make the additional assumption that $\{P_n\}$ is relatively compact. Then, each subsequence $\{P_{n_i}\}$ has a further subsequence $\{P_{n_{i_m}}\}$ that converges weakly to some Borel probability measure Q.

The continuous mapping theorem implies that $P_{n_{i_m}}\pi^{-1}_{t_1,\ldots,t_k} \Rightarrow Q\pi^{-1}_{t_1,\ldots,t_k}$. From our hypothesis we also have $P_n\pi^{-1}_{t_1,\ldots,t_k} \Rightarrow P\pi^{-1}_{t_1,\ldots,t_k}$, so from the uniqueness of the limit, it follows that the finite dimensional distributions of P and Q are identical. However, we saw in example 1.1.17 that the class $\mathcal{C}_f$ of finite dimensional sets is a separating class, whence it follows that $P = Q$. Thus, each subsequence of $\{P_n\}$ has a further subsequence converging weakly to P. According to theorem 1.1.5, this implies that $P_n \Rightarrow P$.

Since C is separable and complete, we can use the above argument together with theorem 1.1.15 to obtain the following result:

Theorem 1.2.3. Let $P, \{P_n\}_{n\geq 1}$ be Borel probability measures on C. If the finite dimensional distributions of P_n converge weakly to those of P, and if $\{P_n\}$ is tight, then $P_n \Rightarrow P$.

This theorem provides a very powerful method for proving weak convergence in C. In order to use it, we should first get a closer look on tightness (and hence on compactness) in this space.

[20] If X is a set and $x \in X$, then the Dirac measure δ_x is defined in the power set $\mathcal{P}(X)$ as follows: $\delta_x(A) = 0$ if $x \notin A$, and $\delta_x(A) = 1$ if $x \in A$.

Recall that, a subset A of a metric space T is called *relatively compact* if $\overline{A}$ is compact.[21] In other words, A is relatively compact if and only if each sequence in A contains a convergent subsequence, whose limit need not belong to A. A very significant result that characterizes relative compacntess in C is the *Arzelà-Ascoli theorem*. To state it, we first need to define the *modulus of continuity* of a function.

Definition 1.2.4. Let $x : [0,1] \to \mathbb{R}$ be an arbitrary function. The *modulus of continuity* of x is a function $w_x : (0,1] \to \mathbb{R}$, defined by

$$w_x(\delta) = \sup_{|s-t|\leq\delta} |x(s) - x(t)|.$$

The modulus of continuity will occasionally be denoted by $w(x,\delta)$ instead of $w_x(\delta)$. Its importance lies on the fact that characterizes continuity in the following manner:

Lemma 1.2.5. A function $x : [0,1] \to \mathbb{R}$ is (uniformly) continuous[22] if and only if

$$\lim_{\delta\to 0^+} w_x(\delta) = 0. \tag{1.12}$$

Proof. We will prove each direction separately:

$(\Rightarrow)$: Suppose that x is continuous. Since x is defined on a closed interval, it is uniformly continuous. Now, consider $\varepsilon > 0$. From uniform continuity, we know that there exists $r > 0$ such that $|x(s) - x(t)| < \varepsilon$ whenever $|s - t| \leq r$. It follows that, for $\delta \in (0, r)$ we have

$$w_x(\delta) = \sup_{|s-t|\leq\delta} |x(s) - x(t)| \leq \sup_{|s-t|\leq r} |x(s) - x(t)| < \varepsilon.$$

Hence, $w_x(\delta)$ tends to zero as $\delta \to 0^+$.

$(\Leftarrow)$: Suppose that $\lim_{\delta\to 0^+} w_x(\delta) = 0$, and consider an arbitrary $\varepsilon > 0$. From our hypothesis, there exists $r \in (0,1)$ such that $w_x(\delta) < \varepsilon$ for $\delta \in (0, r]$. In particular,

$$\sup_{|s-t|\leq r} |x(s) - x(t)| < \varepsilon,$$

which shows that $|x(s) - x(t)| < \varepsilon$ for all $s, t \in [0,1]$ with $|s - t| < r$. Thus, x is uniformly continuous.

□

[21] This notion should not be confused with the notion of relative compactness of a family of measures.
[22] A continuous function defined on a closed interval is automatically uniformly continuous.

From the above lemma we get that the elements of C satisfy (1.12). Another useful result regarding the modulus of continuity is the following one:

Lemma 1.2.6. If $x, y : [0,1] \to \mathbb{R}$ are two continuous functions, then

$$|w_x(\delta) - w_y(\delta)| \leq 2\rho(x,y)$$

for all $\delta \in (0,1]$. So, for fixed $\delta \in (0,1]$, the function $w(x,\delta)$ is (Lipschitz) continuous in x.

Proof. We are only going to show that $w_x(\delta) - w_y(\delta) \leq 2\rho(x,y)$. In a similar way, it can be shown that $w_y(\delta) - w_x(\delta) \leq 2\rho(y,x) = 2\rho(x,y)$, which, combined with the above relation, give the desired inequality. We consider an arbitrary $\varepsilon > 0$. From the definition of $w_x(\delta)$, we can choose $s, t \in [0,1]$ with $|s-t| \leq \delta$, such that

$$|x(s) - x(t)| + \varepsilon > w_x(\delta). \tag{1.13}$$

It is also true that $|y(s) - y(t)| \leq w_y(\delta)$, so

$$-|y(s) - y(t)| \geq -w_y(\delta). \tag{1.14}$$

Adding (1.13),(1.14) gives

$$|x(s) - x(t)| - |y(s) - y(t)| + \varepsilon > w_x(\delta) - w_y(\delta).$$

Using the triangle inequality[23] gives that

$$|(x(s) - x(t)) - (y(s) - y(t))| + \varepsilon > w_x(\delta) - w_y(\delta),$$

which is equivalent to

$$|(x(s) - y(s)) - (x(t) - y(t))| + \varepsilon > w_x(\delta) - w_y(\delta).$$

Using the triangle inequality again, we get

$$|x(s) - y(s)| + |x(t) - y(t)| + \varepsilon > w_x(\delta) - w_y(\delta),$$

so, from the definition of ρ it now follows that $2\rho(x,y) + \varepsilon > w_x(\delta) - w_y(\delta)$. Since $\varepsilon > 0$ was arbitrary, we get that $2\rho(x,y) \geq w_x(\delta) - w_y(\delta)$, which completes the proof. □

[23] In the form $|a-b| \geq ||a| - |b||$.

If $A \subset C$ and $t_0 \in [0,1]$, then the functions in A are called *equicontinuous at* t_0 if

$$\lim_{t \to t_0} \sup_{x \in A} |x(t) - x(t_0)| = 0.$$

Also, they are called *uniformly equicontinuous over* $[0,1]$ if

$$\lim_{\delta \to 0} \sup_{x \in A} w_x(\delta) = 0.$$

Notice that uniform equicontinuity implies equicontinuity at all points of $[0,1]$.

Theorem 1.2.7 (Arzelà-Ascoli)**.** A set $A \subset C$ is relatively compact if and only if

$$\sup_{x \in A} |x(0)| < \infty \tag{1.15}$$

and

$$\lim_{\delta \to 0} \sup_{x \in A} w_x(\delta) = 0. \tag{1.16}$$

Proof. Suppose that A is relatively compact. Then, $\overline{A}$ is compact, so it is bounded. It follows that $\sup_{x \in A} \|x\|_\infty < \infty$ so $\sup_{x \in A} |x(0)| < \infty$. We know from lemma 1.2.6 that $w(x, n^{-1})$ is continuous in x. It is also easy to see that it is nonincreasing in n. From lemma 1.2.5 we get that $w(x, n^{-1}) \searrow 0$ for all $x \in A$ as $n \to \infty$. Since $\overline{A}$ is compact, it follows that the convergence is uniform on $\overline{A}$ (so it is also uniform on A). In other words,

$$\lim_{n \to \infty} \sup_{x \in A} w(x, n^{-1}) = 0.$$

Since $w(x, \delta)$ is nonincreasing with respect to δ, it follows from the last relation that

$$\lim_{\delta \to 0} \sup_{x \in A} w_x(\delta) = 0,$$

which is exactly what we wanted to show.

For the converse, assume that both (1.15) and (1.16) hold. From (1.16), we can choose an integer $k \geq 1$ (large enough) so that $\sup_{x \in A} w_x(k^{-1}) < \infty$. Since

$$|x(t)| \leq |x(0)| + \sum_{i=1}^{k} \left| x\left(\frac{it}{k}\right) - \left(\frac{(i-1)t}{k}\right) \right| \leq |x(0)| + k \cdot w_x(k^{-1}),$$

it follows that

$$\sup_{t \in [0,1]} \sup_{x \in A} |x(t)| < \infty. \tag{1.17}$$

We are going to use (1.15) and (1.17) to show that A is totally bounded. This yields that $\overline{A}$ is also totally bounded, so, since C is compact, we will get that $\overline{A}$ is compact (closed and totally bounded).

Suppose that $\sup_{t\in[0,1]} \sup_{x\in A} |x(t)| = M < \infty.$, and consider an arbitrary $\varepsilon > 0$. We choose points $-M = t_0 < t_1 < \ldots < t_k = M$ such that $t_{i+1} - t_i < \varepsilon/2$ for $i = 0, 1, \ldots, k-1$, and let H be the set containing all these points. From (1.16), there exists an integer $k \geq 1$ such that $w_x(k^{-1}) < \varepsilon/2$ for all $x \in A$. Also, let B be the subset of C consisting of the *polygonal functions* that are linear on each interval $I_{ki} = [(i-1)/k, i/k]$, $1 \leq i \leq k$ and take values in H at the endpoints. If $x \in A$ then $|x(i/k)| \leq M$, $1 \leq i \leq k$, hence there exists a point $y \in B$ such that $|x(i/k) - y(i/k)| < \varepsilon/2$, $1 \leq i \leq k$.[24]

Due to the relation $w_x(k^{-1}) < \varepsilon/2$, it follows that $|y(i/k) - x(t)| < \varepsilon$ for all $x \in I_{ki}$. Similarly, $|y((i-1)/k) - x(t)| \leq |y((i-1)/k) - x((i-1)/k)| + |x((i-1)/k) - x(t)| < \varepsilon$ for all $t \in I_{ki}$. Since $y(t)$ is a convex combination of $y((i-1)/k)$ and $y(i/k)$, it follows that $|x(t) - y(t)| < \varepsilon$ for all $t \in I_{ki}$. Notice that i in the above argument can take all the values in $\{1, \ldots, k-1\}$, hence it follows that $|x(t) - y(t)| < \varepsilon$ for all $t \in [0,1]$, which translates into $\rho(x,y) < \varepsilon$. This implies that A can be covered by the (finitely many) balls with radii equal to ε and centers in B. As $\varepsilon > 0$ was chosen arbitrarily, we get that A is totally bounded, which completes the proof. □

Next, we are going to prove a series of intermediate results that will help us prove the main theorem of this section, Donsker's theorem. The statements of these results might look a bit too complicated at first glance, but their proofs are relatively simple.

Theorem 1.2.8. Let $\{P_n\}_{n\geq 1}$ be a sequence of Borel probability measures on C. Then, $\{P_n\}$ is tight if and only if the following two conditions hold:

(i) For each $\eta > 0$, there exist $a > 0$ and and integer $n_0 \geq 1$ such that

$$P_n\left(\{x \in C : \ |x(0)| \geq a\}\right) \leq \eta \text{ for all } n \geq n_0, \tag{1.18}$$

(ii) For each positive ε and η, there exist $\delta \in (0,1)$ and an integer $n_0 \geq 1$ such that

$$P_n\left(\{x \in C : \ w_x(\delta) \geq \varepsilon\}\right) \leq \eta \text{ for all } n \geq n_0. \tag{1.19}$$

Before moving on with the proof of this theorem, we should make a critical remark to clarify the statement and make sure that it has no gaps. More specifically, the statement only makes sense if

[24] This follows from the construction of H; for each point t in $[-M, M]$ there exists a point in H whose distance from t is less than $\varepsilon/2$. Thus, it is possible to construct such a function y.

the sets $\{x \in C : |x(0)| \geq a\}$ and $\{x \in C : w_x(\delta) \geq \varepsilon\}$ belong to the Borel $\sigma-$field of C. For the first one, notice that the projection $\pi_0 : C \to \mathbb{R}$ with $x \mapsto x(0)$ is continuous, and this set is equal to $\pi_0^{-1}((-\infty, -a] \cup [a, +\infty))$. For the second set, notice that, due to lemma 1.2.6, the function $w(\cdot, \delta)$ is continuous in its first argument (for δ fixed) and the set is equal to the inverse image of $[\varepsilon, +\infty)$ under this function.

Also, notice that the second condition can be written in the following form:

(ii)′ For each $\varepsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P_n(\{x \in C : w_x(\delta) \geq \varepsilon\}) = 0. \tag{1.20}$$

Proof of the theorem. Suppose that $\{P_n\}$ is tight. If $\eta > 0$, there exists a compact set K such that $P_n(K) > 1 - \eta$ for all $n \geq 1$. In particular, K is relatively compact, so, from the Arzelà-Ascoli theorem it follows that $K \subset \{x \in C : |x(0)| \leq a\}$ for large a, and $K \subset \{x \in C : w_x(\delta) \leq \varepsilon\}$ for small enough δ. Thus, $\{x \in C : |x(0)| \geq a\} \subset K^c$ for large a, and $\{x \in C : w_x(\delta) \geq \varepsilon\} \subset K^c$ for small δ. Hence, for if a is chosen large enough and δ small enough, then for all $n \geq 1$ we have $P_n(\{x \in C : |x(0)| \geq a\}) \leq P_n(K^c) < \eta$ and $P_n(\{x \in C : w_x(\delta) \geq \varepsilon\}) \leq P_n(K^c) < \eta$. Thus, (i) and (ii) hold.

For the converse implication, we assume that (i) and (ii) hold. Since C is separable and complete, each Borel probability measure P on C is tight (theorem 1.1.10). Thus, from what we have already shown, we get that for $\eta > 0$ there exists $a > 0$ such that $P(\{x \in C : |x(0)| \geq a\}) \leq \eta$, and for $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that $P(\{x \in C : w_x(\delta) \geq \varepsilon\}) \leq \eta$. Therefore, if $\{P_n\}$ satisfies (i) and (ii), we may assume that (1.18) and (1.19) hold also for $n < n_0$, possibly for a larger value of a and a smaller value of δ.[25] Thus, we may assume that $n_0 = 1$.

Given $\eta > 0$, we can choose $a > 0$ such that, if $B = \{x \in C : |x(0)| \leq a\}$ then $P_n(B) \geq 1 - \eta$ for all n (due to condition (i)). Then, we choose $\delta_k > 0$ such that, if $B_k = \{x \in C : w_x(\delta_k) < 1/k\}$, then $P_n(B_k) \geq 1 - \eta/2^k$ for all n (due to condition (ii) for $\varepsilon = 1/k$). Then,

$$P_n\left(B^c \cup \left(\bigcup_{k=1}^{\infty} B_k^c\right)\right) \leq \eta + \sum_{k=1}^{\infty} \frac{\eta}{2^k} = 2\eta$$

for all $n \geq 1$, so

$$P_n\left(B \cap \left(\bigcap_{k=1}^{\infty} B_k\right)\right) \geq 1 - 2\eta \text{ for all } n \geq 1.$$

[25] We can indeed do so using the previous argument separately for $P_1, \ldots, P_{n_0-1}$ and adapt the values of a and δ if needed.

If K is the closure of the set $B \cap (\bigcap_{k=1}^{\infty} B_k)$, then from the above relation it follows that $P_n(K) \geq 1 - 2\eta$ for all $n \geq 1$. Since η was arbitrary, it remains to show that K is compact. Notice that K satisfies (1.15) because $K \subset \overline{B} = B$[26], and it satisfies (1.16) because $K \subset \overline{B_k} = \{x \in C : w_x(\delta_k) \leq 1/k\}$ for all $k \geq 1$[27]. From the Arzelà-Ascoli theorem, we get that K is relatively compact. Since K is closed, it follows that it is compact, and the proof is completed. □

Theorem 1.2.9. Suppose that $0 = t_0 < t_1 < \ldots < t_v = 1$ and

$$\min_{1<i<v} (t_i - t_{i-1}) \geq \delta. \tag{1.21}$$

Then, for arbitrary $x \in C$,

$$w_x(\delta) \leq 3 \max_{1\leq i\leq v} \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1})|, \tag{1.22}$$

and, for an arbitrary Borel probability measure P on C,

$$P(\{x \in C : w_x(\delta) \geq 3\varepsilon\}) \leq \sum_{i=1}^{v} P\left(\left\{x \in C : \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1})| \geq \varepsilon\right\}\right). \tag{1.23}$$

Note that (1.21) does not require $t_i - t_{i-1} \geq \delta$ for $i = 1$ and $i = v$. Also, just like in the previous theorem, it can be shown that the sets inside P are Borel measurable (inverse images of Borel sets under continuous functions), so the statement of the theorem indeed makes sense.

Proof of the theorem. Suppose that

$$\max_{1\leq i\leq v} \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1})| = m.$$

If s, t lie in the same interval $I_i = [t_{i-1}, t_i]$, then $|x(s) - x(t)| \leq |x(s) - x(t_{i-1})| + |x(t) - x(t_{i-1})| \leq 2m$. If they lie in adjacent intervals I_i and I_{i+1} respectively, then $|x(s) - x(t)| \leq |x(s) - x(t_{i-1})| + |x(t_{i-1}) - x(t_i)| + |x(t_i) - x(t)| \leq 3m$. If $|s - t| \leq \delta<$ then s, t mus lie on the same of those intervals or in adjacent ones. Thus,

$$w_x(\delta) = \sup_{|s-t|\leq\delta} |x(s) - x(t)| \leq 3m,$$

[26] B is closed because it is the inverse image of the closed set $(-\infty, -a] \cup [a, +\infty)$ under the continuous function π_0.

[27] For each $k \geq 1$ and $\delta < \delta_k$, and for any $x \in K$ we have $w_x(\delta) \leq w_x(\delta_k) \leq 1/k$, so $\sup_{x\in K} w_x(\delta) < 1/k$, which implies that $\sup_{x\in K} w_x(\delta)$ tends to zero as $\delta \to 0$.

which proves (1.22). From this inequality, it follows that

$$\begin{aligned} P\left(\{x \in C : w_x(\delta) \geq 3\varepsilon\}\right) &\leq P\left(\left\{x \in C : 3 \max_{1\leq i\leq v} \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1})| \geq 3\varepsilon\right\}\right) \\ &= P\left(\bigcup_{i=1}^{v}\left\{x \in C : \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1})| \geq \varepsilon\right\}\right) \\ &\leq \sum_{i=1}^{v} P\left(\left\{x \in C : \sup_{t_{i-1}\leq s\leq t_i} |x(s) - x(t_{i-1}| \geq \varepsilon\right\}\right), \end{aligned}$$

which is the second inequality. □

Corollary 1.2.10. Condition (ii) of theorem 1.2.8 holds if, for all positive ε and η, there exist $\delta \in (0,1)$ and an integer $n_0 \geq 1$ such that

$$\frac{1}{\delta} P_n\left(\left\{x \in C : \sup_{t\leq s\leq t+\delta} |x(s) - x(t)| \geq \varepsilon\right\}\right) \leq \eta,\ n \geq n_0 \tag{1.24}$$

for all $t \in [0,1]$.

Proof. We take $t_i = i\delta$ for $i < v = \lfloor q/\delta \rfloor$. Then, the requirements of theorem 1.2.9 are satisfied. If (1.24) holds, then by (1.23) we get that

$$P\left(\{x \in C : w_x(\delta) \geq 3\varepsilon\}\right) \leq v \cdot \delta\eta,\ n \geq n_0,$$

at which point we applied (1.24) for $t = t_0, t_1, \ldots, t_{v-1}$. Since $v = \lfloor 1/\delta \rfloor \leq 1/\delta$, it follows that $v \cdot \delta\eta \leq \eta$, which shows that condition (ii) of theorem 1.2.8 holds[28]. □

1.2.2 Random Functions

At the introduction of the section we said that our intention is to construe stochastic processes as random elements of C. However, we saw that, for this to happen, we need to prove some measurability conditions for the stochastic processes we want to deal with[29]. In this paragraph we will discuss a method that helps us prove these conditions.

Let $(\Omega, \mathcal{A}, P)$ be a probabiliy space, and let X be a map of Ω into C[30]. This means that $X(\omega) \in C$ with value $X_t(\omega) = X(t,\omega)$ at $t \in [0,1]$. For $t \in [0,1]$ (fixed), let $X_t = X(t) : \Omega \to \mathbb{R}$ be the function whose value at $\omega \in \Omega$ is $X_t(\omega) = X(t,\omega)$. Notice that $X_t = \pi_t \circ X$, where π_t is the projection at t.

[28] In the place of ε we have 3ε but of course this is not a problem as the two conditions are equivalent.

[29] Because random elements are, by definition, measurable functions between some measurable spaces.

[30] From the discussion in the introduction of the section we see that X is actually a stochastic process with continuous sample paths.

Indeed, $(\pi_t \circ X)(\omega) = \pi_t(X(\omega)) = X(t,\omega)$. Also, for $0 \le t_1 < \ldots < t_k \le 1$, let $(X_{t_1}, \ldots, X_{t_k})$ denote the mapping $\Omega \to \mathbb{R}^k$ that sends ω to $(X_{t_1}(\omega), \ldots, X_{t_k}(\omega)) = \pi_{t_1,\ldots,t_k}(X(\omega))$.

If X is a *random function*, that is, a random element of C[31], then, since $\pi_{t_1,\ldots,t_k}$ is (due to continuity) $\mathcal{B}(C)/\mathcal{B}(\mathbb{R}^k)-$measurable, the composition $\pi_{t_1,\ldots,t_k} \circ X$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^k)-$measurable, so $(X_{t_1}, \ldots, X_{t_k}) : \Omega \to \mathbb{R}^k$ is a random vector. In particular, each mapping $X_t : \Omega \to \mathbb{R}$ is a random variable.

Conversely, suppose that X_t is a random variable for all $t \in [0,1]$. It follows that $(X_{t_1}, \ldots, X_{t_k}) = \pi_{t_1,\ldots,t_k} \circ X$ is a random vector for all $0 \le t_1 < t_2, \ldots, t_k \le 1$. Let $A = \pi^{-1}_{t_1,\ldots,t_k}(H)$, $H \in \mathcal{B}(\mathbb{R}^k)$ be a finite dimensional set. Then, $X^{-1}(A) = X^{-1}\pi^{-1}_{t_1,\ldots,t_k}(H) = (\pi_{t_1,\ldots,t_k} \circ X)^{-1}(H)$. Since $\pi_{t_1,\ldots,t_k} \circ X$ is $\mathcal{A}/\mathcal{B}(\mathbb{R}^k)-$measurable, it follows that $X^{-1}(A) \in \mathcal{A}$. In other words, if $\mathcal{C}_f$ is the class of finite dimensional sets, then $X^{-1}(\mathcal{C}_f) \subset \mathcal{A}$. As we have seen, $\mathcal{C}_f$ generates $\mathcal{B}(C)$, so from the previous relation it follows that $X^{-1}(\mathcal{B}(C)) \subset \mathcal{A}$, or, equivalently, that X is $\mathcal{A}/\mathcal{B}(C)-$measurable (that is, X is a random function). Thus, we have shown that X is a random function if and only if X_t is a random variable for all $t \in [0,1]$.

Finally, if $P_X = PX^{-1}$ is the distribution of X, then for the finite dimensional distributions of P_X we have

$$\begin{aligned} P_X\pi^{-1}_{t_1,\ldots,t_k}(H) &= P_X\left(\pi^{-1}_{t_1,\ldots,t_k}(H)\right) \\ &= P\left(X^{-1}(\pi^{-1}_{t_1,\ldots,t_k}(H))\right) \\ &= P\left((\pi_{t_1,\ldots,t_k} \circ X)^{-1}(H)\right) \\ &= P\left((X_{t_1}, \ldots, X_{t_k}) \in H\right). \end{aligned}$$

Thus, the finite dimensional distributions of P_X are the distributions of the corresponding random vectors $(X_{t_1}, \ldots, X_{t_k})$.

Let $X, X^1, X^2, \ldots$ be random functions.

Theorem 1.2.11. If

$$(X^n_{t_1}, \ldots, X^n_{t_k}) \Rightarrow (X_{t_1}, \ldots, X_{t_k}) \tag{1.25}$$

is true for all $t_1, \ldots, t_k$ and if

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\left(\{w(X^n, \delta) \ge \varepsilon\}\right) = 0 \tag{1.26}$$

for each positive ε, then $X^n \Rightarrow X$.

[31] Which in turn means that X is $\mathcal{A}/\mathcal{B}(C)-$measurable.

Proof. From our hypothesis and the above discussion it follows that (1.26) is equivalent to $P_{X^n}\pi^{-1}_{t_1,\dots,t_k} \Rightarrow P_X\pi^{-1}_{t_1,\dots,t_k}$ for all $t_1,\dots,t_k$. Also, $X^n \Rightarrow X$ is equivalent to $P_{X^n} \Rightarrow P_X$. Thus, the result will follow from theorem 1.2.3 if we also show that $\{P_{X^n}\}$ is tight. We know that $X_0^n \Rightarrow X_0$,so $\{P_{X^n}\pi_0^{-1}\}$ is tight. Indeed, this is the distribution of X_0^n, so it converges weakly to the distribution $P_X\pi_0^{-1}$ of X_0. Since $\{P_{X^n}\pi_0^{-1}\}$ is weakly convergent, it follows that it is relatively compact, so, from the separability and completeness of C and from Prohorov's theorem (1.1.15) we get that $\{P_{X^n}\pi_0^{-1}\}$ is tight. Thus, for each $\eta > 0$, there exists a compact set $K \subset \mathbb{R}$ such that $P_{X^n}\pi_0^{-1}(K) > 1-\eta$ for all $n \geq 1$, which is exactly condition (i) of the theorem 1.2.8. We will show that the second condition is also satisfied. We have

$$\begin{aligned} P\left(w(X^n(\omega),\delta) \geq \varepsilon\right) &= P\left(X^n(\omega) \in \{x \in C : w(x,\delta) \geq \varepsilon\}\right) \\ &= P\left(\left(X^n\right)^{-1}\left(\{x \in C : w(x,\delta) \geq \varepsilon\}\right)\right) \\ &= P_{X^n}\left(\{x \in C : w(x,\delta) \geq \varepsilon\}\right) \end{aligned}$$

so from the hypothesis it follows that

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P_{X^n}\left(\{x \in C : w(x,\delta) \geq \varepsilon\}\right) = 0,$$

which is exactly condition (ii)' of the theorem 1.2.8. Since both conditions of this theorem are satisfied, it follows that $\{P_{X^n}\}$ is tight. □

We have already explained that the projection $\pi_t : C \to \mathbb{R}$ is Borel measurable, so it can be viewed as a random variable defined on C. We will denote this random variable by x_t. Thus, for a fixed t, the random variable x_t has value $x(t)$ at x. If P is a Borel probability measure on C and t is thought of as a time parameter, then $\{x_t\}_{0\leq t\leq 1}$ becomes a stochastic process, and the x_t are commonly called the *coordinate variables*. We can also consider the distribution of x_t under P, defined as $P_{x_t}(H) = P(\{x \in C : x_t \in H\})$ and often written as $P\left(x_t \in H\right)$.

1.2.3 Wiener's Measure and Donsker's Theorem

The first purpose of this chapter is to prove the existence of the *Wiener process*, a random function whose properties are stated in section 1. Our method will be the following: we shall first prove the existence of the *Wiener measure*, a Borel probability measure on C having certain special properties, and then we will construct the Wiener process as a random function whose distribution will be the Wiener measure.

The Wiener measure is a Borel probability measure on C with the following properties:

- Each coordinate variable x_t is normally distributed with mean zero and variance t. More specifically,
$$W\left(\{x \in C : x(t) \leq a\}\right) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{a} e^{-x^2/2t}\, dx \text{ for } t > 0. \tag{1.27}$$
By this definition, the coordinate variable x_0 is almost surely equal to zero under W.

- The stochastic process $Y = Y(x,t)$ with $Y(x,t) = x(t)$ [32] has independent increments under W. That is, if $0 = t_0 \leq t_1 \leq \ldots \leq t_k \leq 1$, then the coordinate random variables
$$x_{t_1} - x_{t_0}, x_{t_2} - x_{t_1}, \ldots, x_{t_k} - x_{t_{k-1}} \tag{1.28}$$
are independent under W [33].

We are first going to show that there exists at most one such measure. In example 1.1.17 we saw that each measure on $(C, \mathcal{B}(C))$ is uniquely determined by its finite dimensional distributions, so it suffices to show that the above two properties uniquely determine the finite dimensional distributions of W.

If $s \leq t$ then the random variable $x_t \sim \mathcal{N}(0,t)$ is the sum of the independent random variables $x_s - x_0 = x_s \simeq \mathcal{N}(0,s)$ and $x_t - x_s$. It follows that $x_t - x_s \sim \mathcal{N}(0, t-s)$, which shows that the stochastic process Y defined above has stationary increments [34]. Thus, if $0 = t_0 \leq t_1 \leq \ldots \leq t_k \leq 1$ and $a_1, \ldots, a_k \in \mathbb{R}$, it follows from (1.28) that

$$W\left(x_{t_i} - x_{t_{i-1}} \leq a_i,\ i = 1, \ldots, k\right) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \int_{-\infty}^{a_i} e^{-x^2/2(t_i - t_{i-1})}\, dx. \tag{1.29}$$

We observe that $x_s x_t = x_s^2 + x_s(x_t - x_s)$ so, using the independence of x_s and $x_t - x_s$, we get that

$$\begin{aligned}
\text{Cov}(x_s, x_t) &= \mathbb{E}(x_s x_t) - \mathbb{E}(x_s)\mathbb{E}(x_t) \\
&= \mathbb{E}(x_s^2) + \mathbb{E}(x_s(x_t - x_s)) \\
&= \text{Var}(x_s) + \mathbb{E}(x_s)^2 + \mathbb{E}(x_s)\mathbb{E}(x_t - x_s) \\
&= s.
\end{aligned}$$

Due to the second property of the Wiener measure, equation (1.29) completely characterizes the distribution of the random vector $v = (x_{t_1}, x_{t_2} - x_{t_1}, \ldots, x_{t_k} - x_{t_{k-1}})^T$ [35]. In particular, this vector follows

[32] Y is defined on the probability space $(C, \mathcal{B}(C), W)$ and is indexed by $[0,1]$.
[33] Note that these random variables are defined on $(C, \mathcal{B}(C), W)$.
[34] That is, the distribution of $x_t - x_s$ for $t \geq s$ depends only on $t - s$.
[35] We considered the transpose simply because we want to perform some matrix operations later.

the $k-$variate normal distribution with zero mean vector and covariance matrix

$$\begin{pmatrix} t_1 & 0 & \dots & 0 \\ 0 & t_2 - t_1 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & t_k - t_{k-1} \end{pmatrix}.$$

If A is the matrix

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix},$$

then $Av = (x_{t_1}, \dots, x_{t_k})^T$, so the vector $(x_{t_1}, \dots, x_{t_k})$ is also normally distributed [36]. From the above discussion, it follows that its mean vector is the zero vector, and its covariance matrix has the element $\min(t_i, t_j)$ is the position (i, j).

We deduce that the distribution of $(x_{t_1}, \dots, x_{t_k})^T$ is uniquely determined. By extension, the finite dimensional distributions of W are completely characterized by its two defining properties. As we explained, this implies that there exists at most one choice for W.

Wiener's measure will come up as the weak limit of a sequence of probability measures. First, let $\xi_1, \xi_2, \dots : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be a sequence of i.i.d. random variables with $\mathbb{E}(\xi_i) = 0$ and $\mathrm{Var}(\xi_i) = \sigma^2 > 0$. We consider the partial sums $S_n = \xi_1 + \dots + \xi_n$ of the sequence, with the convention $S_0 = 0$, and let $X^n(\omega)$ be a random element of C constructed as follows: for $i = 0, 2, \dots, n$, $X^n(\omega)$ has the value $S_i(\omega)/\sigma\sqrt{n}$ at $t = i/n$. Then, $X^n(\omega)$ is extended linearly on each subinterval $[(i-1)/n, i/n]$. In other words,

$$X_t^n(\omega) = \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor}(\omega) + \frac{1}{\sigma\sqrt{n}} (nt - \lfloor nt \rfloor) \xi_{\lfloor nt \rfloor + 1}(\omega), \ t \in [0, 1]. \tag{1.30}$$

The random element X^n is a random function, whose underlying probability space is Ω.

We consider the random variable $\psi_{n,t} = \frac{1}{\sigma\sqrt{n}} (nt - \lfloor nt \rfloor) \xi_{\lfloor nt \rfloor + 1}$, which appears at (1.30). We have $\mathbb{E}(\psi_{n,t}) = 0$ because $\mathbb{E}(\xi_i) = 0$ for all i. Also,

$$\mathrm{Var}(\psi_{n,t}) = \frac{(nt - \lfloor nt \rfloor)^2}{n\sigma^2} \mathbb{E}(\xi^2_{\lfloor nt \rfloor + 1}) = \frac{(nt - \lfloor nt \rfloor)^2}{n\sigma^2} \mathrm{Var}(\xi_{\lfloor nt \rfloor + 1}) = \frac{(nt - \lfloor nt \rfloor)^2}{n}.$$

[36] As linear transformation of a normally distributed vector.

We shall show that $\psi_{n,t} \Rightarrow 0$ as $n \to \infty$. For $t = i/n$ for some $i = 0, \ldots, n$ this is obvious as $\psi_{n,t}$ Assume that t is not of this form. Then, for any $\varepsilon > 0$, Chebyshev's inequality gives

$$P\left(|\psi_{n,t}| \geq \varepsilon\right) = P\left(|\psi_{n,t}| \geq \frac{\sqrt{n}\varepsilon}{nt - \lfloor nt \rfloor}\sqrt{\mathrm{Var}(\psi_{n,t})}\right) \leq \frac{(nt - \lfloor nt \rfloor)^2}{n\varepsilon^2} \leq \frac{1}{n\varepsilon^2}.$$

The last quantity obviously tends to 0 as $n \to \infty$ so $P\left(|\psi_{n,t}| \geq \varepsilon\right) \to 0$ as $n \to \infty$. This shows that $\psi_{n,t} \xrightarrow{P} 0$ [37], which, as we know, implies that $\psi_{n,t} \xrightarrow{d} 0$.

We shall now show that $X_t^n \Rightarrow \sqrt{t}N$, where N has the standard normal distribution. First of all, from the Central Limit Theorem it follows that $\frac{S_{\lfloor nt \rfloor}}{\sigma\sqrt{\lfloor nt \rfloor}} \Rightarrow N$. Also, $\sqrt{\frac{\lfloor nt \rfloor}{n}} \to \sqrt{t}$ so

$$\frac{S_{\lfloor nt \rfloor}}{\sigma\sqrt{n}} = \sqrt{\frac{\lfloor nt \rfloor}{n}} \cdot \frac{S_{\lfloor nt \rfloor}}{\sigma\sqrt{\lfloor nt \rfloor}} \Rightarrow \sqrt{t}N.$$

We have $|X_t^n - \psi_{n,t}| \Rightarrow \sqrt{t}N$ and $\psi_{n,t} \Rightarrow 0$, which yields that $X_t^n \Rightarrow \sqrt{t}N$, which proves our claim. Similarly, if $s \leq t$ then

$$\begin{aligned}(X_s^n, X_t^n - X_s^n) &= \frac{1}{\sigma\sqrt{n}}\left(S_{\lfloor ns \rfloor}, S_{\lfloor nt \rfloor} - S_{\lfloor ns \rfloor}\right) + \left(\psi_{n,s}, \psi_{n,t} - \psi_{n,s}\right) \\ &\Rightarrow (N_1, N_2),\end{aligned}$$

where the independent random variables N_1, N_2 are normal, with means equal to zero and variances s and $t - s$ respectively. The independence of N_1, N_2 follows from the independence of $\xi_1, \xi_2, \ldots$ By extension, if $0 \leq t_1 \leq, \ldots \leq t_k \leq 1$, then

$$\left(X_{t_1}^n, X_{t_2}^n - X_{t_1}^n, \ldots, X_{t_k}^n - X_{t_{k-1}}^n\right)^T \Rightarrow \left(N_1, \ldots, N_k\right)^T, \tag{1.31}$$

where the independent random variables $N_1, \ldots, N_k$ are normal, with mean zero and variances $t_1, t_2 - t_1, \ldots, t_k - t_{k-1}$ respectively. Using the continuous mapping theorem for the continuous linear map induced by the matrix

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

gives us that the limiting distribution of $(X_{t_1}^n, \ldots, X_{t_k}^n)^T$ is the same as the one of corresponding finite dimensional distribution of the measure W we want to construct.

[37] Convergence in probability

In other words, if P_{X^n} is the distribution of the random element X^n of C, then $P_{X^n}\pi^{-1}_{t_1,\dots,t_k}$, which is the distribution of $(X^n_{t_1},\dots,X^n_{t_k})^T$ on $\mathbb{R}^k$, converges weakly to what we want $W\pi^{-1}_{t_1,\dots,t_k}$ to be. For now, we will denote the limiting distribution of $P_{X^n}\pi^{-1}_{t_1,\dots,t_k}$ by $\mu_{t_1,\dots,t_k}$.

Our purpose is to show that $\{P_{X^n}\}_{n\geq 1}$ is tight. In this case, Prohorov's theorem yields that there exists some subsequence $\{P_{X^{k_n}}\}_{n\geq 1}$ converging weakly to a Borel probability measure U on C. Then, $P_{X^{k_n}}\pi^{-1}_{t_1,\dots,t_k} \Rightarrow U\pi^{-1}_{t_1,\dots,t_k}$ so $U\pi^{-1}_{t_1,\dots,t_k} = \mu_{t_1,\dots,t_k}$ for all $t_1,\dots,t_k \in [0,1]$. This essentially means that the finite dimensional distributions of U are $\mu_{t_1,\dots,t_k}$, which are equal to what we want $W\pi^{-1}_{t_1,\dots,t_k}$ to be. Since the finite dimensional distributions characterize a measure, we deduce that U is exactly the Wiener measure whose existence we are trying to prove.

The tightness of $\{P_{X^n}\}_{n\geq 1}$ will follow more naturally if we first examine the more general case when $\{\xi_i\}_{i\geq 1}$ is stationary, that is, the distribution of the vector $(\xi_k,\dots,\xi_{k+i})$ depends only on its length, and is independent of k. Obviously, the sequence $\{\xi_i\}_{i\geq 1}$ we considered before has this property. The next lemma is very technical but, after we prove it, the existence of the Wiener measure will follow very easily.

Lemma 1.2.12. Suppose that $X_n : (\Omega_n, \mathcal{A}_n, P_n) \to C[0,1]$ is the sequence of random elements of $C[0,1]$ defined by (1.30), and that $\xi_i : (\Omega, \mathcal{A}, P) \to \mathbb{R}$, $i \geq 1$ is stationary. If

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \lambda^2 P\left(\max_{k\leq n} |S_k| \geq \lambda\sigma\sqrt{n}\right) = 0, \tag{1.32}$$

then $\{P_{X^n}\}$ is tight.

Proof. We shall use theorem 1.2.8. Since $X^n_0 = 0$ for all $n \geq 1$, its first condition is trivially satisfied. As we have explained, the second condition of that theorem is equivalent to (1.20). In our case, this relation is equivalent to the requirement that

$$\lim_{\delta\to 0} \limsup_{n\to\infty} P_n\left[w(X^n,\delta) \geq \varepsilon\right] = 0, \text{ for all } \varepsilon > 0. \tag{1.33}$$

This relation can be shown using theorem 1.2.9. According to this theorem, if (1.21) holds, then (1.22) and (1.23) are also true, so

$$P_n\left(w\left(X^n,\delta\right) \geq 3\varepsilon\right) \leq \sum_{i=1}^{v} P_n\left(\sup_{t_{i-1}\leq s<t_i} \left|X^n_s - X^n_{t_{i-1}}\right| \geq \varepsilon\right) \tag{1.34}$$

if $\min_{1<i<v}(t_i - t_{i-1}) \geq \delta$. This is easier to analyze if we choose $t_i = m_i/n$, where $m_0, m_1, \dots, m_v$ are integers satisfying $0 = m_0 < m_1 < \dots < m_v = n$. The reason for this specific choice of the points

$t_0, \ldots, t_v$ is that the functions $X^n(\omega)$ are polygonal for all $\omega \in \Omega_n$, so the supremum in (1.35) becomes a maximum of differences $\left|S_k - S_{m_{i-1}}\right| / \sigma\sqrt{n}$, where $m_{i-1} \leq k \leq m_i$. Indeed, since $X_s^n(\omega)$, $s \in [t_{i-1}, t_i]$ is a polygonal function, its maximum difference from $X^n_{m_{i-1}/n} = S_{m_{i-1}}/\sigma\sqrt{n}$ as s varies, is achieved in one of the points of the set $\left\{\frac{k}{n} : k = 0, 1, \ldots, v\right\} \cap [t_{i-1}, t_i]$. Since $X^n_{k/n} = S_k/\sigma\sqrt{n}$, it follows that

$$\sup_{t_{i-1} \leq s \leq t_i} \left|X_s^n - X^n_{t_{i-1}}\right| = \max_{m_{i-1} \leq k \leq m_i} \frac{1}{\sigma\sqrt{n}} \left|S_k - S_{m_{i-1}}\right|.$$

Since $\{\xi_i\}_{i \geq 1}$ is stationary, it follows that the latter maximum has the same distribution as

$$\max_{0 \leq k \leq m_i - m_{i-1}} |S_k|.$$

Hence, from (1.34) it follows that

$$P_n\left(w(X^n, \delta) \geq 3\varepsilon\right) \leq \sum_{i=1}^{v} P_n\left(\max_{0 \leq k \leq m_i - m_{i-1}} \frac{|S_k|}{\sigma\sqrt{n}} \geq \varepsilon.\right). \tag{1.35}$$

Recall that this inequality only holds if $t_i - t_{i-1} \geq \delta$ for all $i \in \{2, \ldots, v-1\}$. This translates as $m_i - m_{i-1} \geq n\delta$ for all $i \in \{2, \ldots, i-1\}$. To simplify our analysis even more, we shall choose $m_i = im$ for $0 \leq i < v$, and $m_v = n$, where m is an integer. Since we want $m_i - m_{i-1} \geq n\delta$, we must have $m \geq n\delta$, so we can choose $m = \lceil n\delta \rceil$. We also need $m_{v-1} = (v-1)m \leq n \leq vm$ so $v \geq n/m$ and $v - 1 \leq n/m$. Thus, we can set $v = \lceil n/m \rceil$. Notice that $n/m = n/\lceil n\delta \rceil$ converges to $1/\delta$ as $n \to \infty$, due to the squeezing theorem. Thus, for n large enough and δ small enough, we have $n/m > 1/2\delta$. Also, $v = \lceil n/m \rceil$ converges to $\lceil 1/\delta \rceil$ which is smaller than $1 + 1/\delta < 2/\delta$ for small enough δ. Since $m_v - m_{v-1} \leq m$, it follows from (1.35) that, for large n and small δ we have

$$\begin{aligned} P_n\left(w(X^n, \delta) \geq 3\varepsilon\right) &\leq v \cdot P_n\left(\max_{0 \leq k \leq m} |S_k| \geq \varepsilon\sigma\sqrt{n}\right) \\ &\leq \frac{2}{\delta} \cdot P_n\left(\max_{0 \leq k \leq m} |S_k| \geq \frac{\varepsilon}{\sqrt{2\delta}}\sigma\sqrt{m}\right) \end{aligned}$$

where the last inequality follows directly from the fact that $n/m > 1/2\delta$. We can now choose $\lambda = \varepsilon/\sqrt{2\delta}$ in which case the above inequality translates into

$$P_n\left(w(X^n, \delta) \geq 3\varepsilon\right) \leq \frac{4\lambda^2}{\varepsilon^2} \cdot P_n\left(\max_{0 \leq k \leq m} |S_k| \geq \lambda\sigma\sqrt{m}\right).$$

Given positive ε, η, (1.34) implies that there exists some λ such that

$$\frac{4\lambda^2}{\varepsilon^2} \limsup_{m \to \infty} P_n\left(\max_{0 \leq k \leq m} |S_k| \geq \lambda\sigma\sqrt{m}\right) < \eta.$$

Notice that $\delta \to 0$ as $\lambda \to \infty$, and, for fixed δ, $m \to \infty$ as $n \to \infty$. Thus, from the above relations it follows easily that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P_n \left(w(X^n, \delta) \geq \varepsilon \right) = 0 \text{ for all } \varepsilon > 0,$$

which completes the proof. □

The construction of the Wiener measure can be completed using the independence of the random variables $\{\xi_i\}_{i \geq 1}$ Because of the independence, $\{\xi_i\}_{i \geq 1}$ is stationary, and Etemadi's inequality gives that

$$P_n \left(\max_{0 \leq u \leq m} |S_u| \geq \alpha \right) \leq 3 \max_{0 \leq u \leq m} P \left(|S_u| \geq \alpha / 3 \right) \text{ for all } \alpha > 0. \tag{1.36}$$

To prove the second condition of lemma 1.2.12, we thus need to show that

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \left\{ \lambda^2 \max_{0 \leq k \leq n} P \left(|S_k| \geq \lambda \sigma \sqrt{n} \right) \right\} = 0. \tag{1.37}$$

The interesting fact is that we can use any sequence $\{\xi_i\}_{i \geq i}$ we want. We assume that each ξ_i is a standard normal random variable. Of course, $\xi_1, \xi_2, \ldots$ are also assumed independent. In this case, we know that $S_k \sim \mathcal{N}(0, k)$ [38] so $S_k / \sqrt{k} \sim \mathcal{N}(0, 1)$. From Markov's inequality it follows that

$$\begin{aligned} P \left(|N| \geq \lambda \right) &= P \left(N^4 \geq \lambda^4 \right) \\ &\leq \frac{\mathbb{E}(N^4)}{\lambda^4} \\ &= \frac{3}{\lambda^4} \end{aligned}$$

as $\mathbb{E}(N^4) = 3$. Thus, $P_n \left(|S_k| \geq \lambda \sigma \sqrt{n} \right) = P_n \left(\sqrt{k} |N| \geq \lambda \sigma \sqrt{n} \right) \leq 3 / \lambda^4 \sigma^4$ for $k \leq n$. Letting $\lambda \to \infty$ proves (1.37), which in turn proves the existence of Wiener's measure.

Theorem 1.2.13. There exists on $(C, \mathcal{B}(C))$ a measure whose finite dimensional distributions are determined by (1.29).

By W, we will denote not only the Wiener measure, but also any random function having the Wiener measure as its distribution over C.

We are now ready to prove Donsker's theorem. Notice that, in the above discussion, W came up as the weak limit of the sequence $\{P_{X^n}\}_{n \geq 1}$, where X^n was defined by (1.30), and $\{\xi_i\}_{i \geq 1}$ are independent *standard normal* random variables. Donsker's theorem actually shows that we can drop the requirement of normality of those variables. It was introduced by [Donsker, 1951].

[38] We can show this result using characteristic functions.

Theorem 1.2.14 (Donsker). If $\xi_1, \xi_2, \ldots$ are independent and identically distributed random variables with mean zero and variance σ^2, and if X^n is the random function defined by (1.30), then $X^n \Rightarrow W$.

Proof. The proof depends on theorem 1.2.11. If $W : (\Omega, \mathcal{A}, P) \to C$ denotes the random function whose distribution over C is the Wiener measure [39]. From the definition of the Wiener measure and the discussion at the beginning of the section, it follows that $W_s \sim \mathcal{N}(0, s)$ for all $s \geq 0$. Indeed,

$$\begin{aligned} P_{W_t}\left(\{x \in C : x(t) \leq a\}\right) &= P\left(W_t \leq a\right) \\ &= P\left(\pi_t \circ W \leq a\right) \\ &= P\left(W \in \pi_t^{-1}\left((-\infty, a]\right)\right) \\ &= W\left(\pi_t^{-1}\left((-\infty, a]\right)\right) \\ &= W\left(\{x \in C : x(t) \leq a\}\right) \\ &= \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{a} e^{-x^2/2t}\, dx, \end{aligned}$$

where, in the last two lines W denotes the Wiener measure, which is the distribution of the random element W. We can similarly show that $W_t - W_s \sim \mathcal{N}(0, t-s)$ for $t \geq s$. In the beginning of the section, we saw that $(X_s^n, X_t^n - X_s^n) \Rightarrow (N_1, N_2)$, where N_1, N_2 are independent random variables with mean zero and variances s and $t - s$ respectively. From the second property of the definition of the Wiener measure, it follows that $W_t - W_s$ and $W_s = W_s - W_0$ are independent so the last relation can be written as $(X_s^n, X_t^n - X_s^n) \Rightarrow (W_s, W_t - W_s)$. Due to the continuous mapping theorem, it follows that $(X_s^n, X_t^n) \Rightarrow (W_s, W_t)$. By extension, $(X_{t_1}^n, \ldots, X_{t_k}^n) \Rightarrow (W_{t_1}, \ldots, W_{t_k})$ for all $0 \leq t_1 \leq \ldots \leq t_k$, which is relation (1.25) with W in the role of X.

It remains to prove (1.26). We have shown this relation in the proof of the previous theorem, but under the additional assumption that ξ_i are normal. To drop this assumption, we shall use the central limit theorem. Notice that, from the proof of the previous theorem, it suffices to prove (1.37). We fix some $\lambda > 0$. By the central limit theorem, if k_λ is large enough, greater than some k_λ, then $S_k/\sigma\sqrt{k} \approx \mathcal{N}(0,1)$. Markov's inequality for a random variable $Z \sim \mathcal{N}(0,1)$ gives $P\left(|Z| \geq \lambda\right) = P\left(Z^4 \geq \lambda^4\right) \leq \mathbb{E}(Z^4)/\lambda^4 = 3/\lambda^4$ so, for large k we have $P\left(|S_k| \geq \lambda\sigma\sqrt{k}\right) < 4/\lambda^4$. For $k \leq n$ it is true that $P\left(|S_k| \geq \lambda\sigma\sqrt{k}\right) \geq P\left(|S_k| \geq \lambda\sigma\sqrt{n}\right)$, whence it follows that

$$\lambda^2 \max_{k \leq n} P\left(|S_k| \geq \lambda\sigma\sqrt{n}\right) < 4/\lambda^2$$

[39] We have explained that, for each measure there exists a random element whose distribution is exactly this measure.

for $k \geq k_\lambda$. Notice that $\text{Var}(S_k) = k\sigma^2$ so, for $k < k_\lambda$, Chebyshev's inequality gives that $P\left(|S_k| \geq \lambda\sigma\sqrt{n}\right) \leq k/\lambda^2 n \leq k_\lambda/n\lambda^2$. Thus, $\max_{k\leq n} P\left(|S_k| \geq \lambda\sigma\sqrt{n}\right) \leq \max\left\{k_\lambda/n\lambda^2, 4/\lambda^4\right\}$, so it obviously tends to zero as $\lambda \to \infty$. This shows that (1.32) is true, so from the proof of the previous theorem we obtain that (1.26) is also true. Hence, theorem 1.2.11 implies that $X^n \Rightarrow W$, which completes the proof. □

Chapter 2

Empirical Process Theory and Applications in Statistics

In the second part of this thesis, we are going to focus on uniform laws of large numbers and uniform central limit theorems. We present these topics within the framework of Empirical Process Theory. Our starting point are the Strong Law of Large Numbers (SLLN).

The SLLN states that, if $Z_1, \dots, Z_n : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{A})$ are independent and identically distributed random elements with distribution P_Z in $(\mathcal{Z}, \mathcal{A})$, and if $f : \mathcal{Z} \to \mathbb{R}$ is a measurable function such that $\mathbb{E}f(Z)$ exists, then

$$\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \overset{n\to\infty}{\longrightarrow} \mathbb{E}f(Z),\ P-\text{a.s.} \tag{2.1}$$

This result is one of the most prominent landmarks in Probability Theory. One of its countless areas of application is consistency of statistical estimators, and, in particular, plug-in estimators. This application is presented in Example 2.0.1. Before moving forward to ths application, we recall the definition of the empirical measure.

Let $Z_1, \dots, Z_n : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{G})$ be i.i.d. random elements with distribution P_Z. The empirical measure $\mathbb{P}_n$ induced by these elements is defined by

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z_i \in A\}} = \frac{\#\{1 \le i \le n : Z_i \in A\}}{n},\ A \in \mathcal{G}. \tag{2.2}$$

Notice that the empirical measure is a random measure in the sense that it depends on the value of $\omega \in \Omega$ we choose each time. If the singletons in $(\mathcal{Z}, \mathcal{G})$ are measurable, then the empirical measure can be described alternatively as the random probability measure that gives mass $1/n$ to each of the

observations Z_i. For this reason, the empirical measure is usually expressed as

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i},$$

where δ_z denotes the Dirac measure at point $z \in \mathcal{Z}$. The SLLN can be expressed via the empirical measure in the equivalent form

$$\mathbb{P}_n f \xrightarrow{a.s.} P_Z f. \tag{2.3}$$

Indeed, the elements $f(Z_1), \ldots, f(Z_n)$ are i.i.d. random variables, and their expectation is equal to

$$\begin{aligned} \mathbb{E}\left(f(Z_i)\right) &= \int_\Omega f(Z_i(\omega))\, dP(\omega) \\ &= \int_T f\, dP_Z \\ &= P_Z f. \end{aligned} \tag{2.4}$$

Combined with the fact that $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$,[1] Equation (2.4) verifies the equivalent form (2.3) of the SLLN.

Example 2.0.1 (Plug-in estimators)**.** Some of the most common quantities of interest in statistics, like the expectation and the quantiles, can be expressed as functionals of the underlying distribution. For example, one can define the expectation of a random variable Y as a functional $\gamma : \mathcal{P} \to \mathbb{R}$ defined by

$$\gamma(Q) = \int_{\mathbb{R}} Y\, dQ, \tag{2.5}$$

where $\mathcal{P}$ is the space of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $Y \sim Q$. This maps any probability measure $Q \in \mathcal{P}$ to $\mathbb{E}[Y]$, where $Y \sim Q$. Analogously, for any fixed threshold $\alpha \in (0,1)$, the α-quantile of the distribution of $g(Y)$ can be expressed as

$$\gamma_\alpha(Q) = \inf\left\{t \in \mathbb{R} : Q\left((-\infty, t]\right) \geq \alpha\right\}. \tag{2.6}$$

Let $\Gamma : \mathcal{P} \to \mathbb{R}$ be any such functional. Then, for any $Q \in \mathcal{P}$, the (random) estimator $\Gamma(Q_n)$ is called a *plug-in estimator* of $\Gamma(Q)$, where Q_n is the empirical measure corresponding to an i.i.d. sequence $Y_1, \ldots, Y_n \sim Q$.

Plug-in estimators are useful because they can be computed using the observed data points $Y_1, Y_2, \ldots$

[1] For each measurable function $f : T \to \mathbb{R}$ it is true that $\delta_z f = f(z)$.This simple property can be shown in a standard way: first for indicator functions, then for simple functions, then for non-negative measurable functions using the Monotone Convergence Theorem, and finally for general measurable functions. Consequently, for every measurable function $f : T \to \mathbb{R}$ we have $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$.

unlike estimators that depend directly on the underlying distribution Q, which remains unknown. An important property that makes plug-in estimators suitable for estimating the target quantity is consistency, that is, convergence to the target quantity, either almost surely (strong consistency), or in probability (weak consistency). For example, the SLLN can be used to yield strong consistency for the plug-in estimator $\gamma(Q_n)$ of the mean, where the functional γ is defined in Equation (2.5).

In a similar way, we can use the SLLN to derive consistency of an abundance of other plug-in estimators. However, rather than expanding our discussion on the use of these two theorems in deriving useful properties of statistical estimators, we will now investigate some of their most restrictive limitations.

2.1 Limitations of the SLLN

Although the SLLN is a rich source of a multitude of methods and techniques in Probability and Statistics, it still faces severe limitations. The main root of these limitations is the assumption that g is a nonrandom function. On the contrary, some of the most established statistical methods make use of functions that depend on the observed data.

In this subsection, we will present some of these statistical methods and explain why these cases lie outside of the area of validity of the SLLN.

Example 2.1.1 (Empirical Risk Minimization)**.** Suppose that $Z_1, \ldots, Z_n : (\Omega, \mathcal{A}, P) \to \mathcal{Z}$ are i.i.d. random elements whose distribution belongs to a class $\{P_f : f \in \mathscr{F}\}$, and is thus determined by an unknown function $f \in \mathscr{F}$. Let $f^\star$ denote the true (unknown) value of the function, which we would like to estimate. The estimation relies on a loss function $f \mapsto \mathcal{L}_f(Z)$, whose expectation with respect to $P_{f^\star}$ we would like to minimize. The loss function is chosen in such a way that the true value $f^\star$ is the minimizer of the expected loss $\mathbb{E}_{f^\star}\left[\mathcal{L}_f(Z)\right]$,[2] which is called the *risk* and is denoted by $R\left(f, f^\star\right)$. The risk is not computable, because $f^\star$ is unknown. Hence, we cannot compute the value of f that minimizes it. Hence, motivated by the LLN, we resort to its empirical counterpart, namely

$$\widehat{R}_n\left(f\right) := \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_f\left(\mathrm{Z}_i\right). \tag{2.7}$$

We can now compute the value $\widehat{f}_n$ that minimizes $\widehat{R}_n\left(f\right)$ and hope that this value is close to the minimizer of $\arg\min_{f\in\mathscr{F}} R\left(f, f^\star\right) = f^\star$. This general procedure is widely known as Empirical Risk Minimization (ERM). We shall look at the following examples of ERM:

[2]The index $f^\star$ denotes that the expectation is taken over the true distribution, which is $P_{f^\star}$. This is totally reasonable, as we want to minimize the true expected loss, and not the loss based on some false value of the parameter.

1. Maximum Likelihood Estimation (MLE): In MLE, we assume that the function class $\mathscr{F}$ is equal to a parameter space Θ, and that the distribution of Z is determined by a parameter $\theta \in \Theta$. Assume that $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ, and we denote by p_θ the density of P_θ with respect to μ. The loss function is defined by

$$\mathcal{L}_\theta(x) = \log \frac{p_{\theta^\star}(x)}{p_\theta(x)}. \tag{2.8}$$

Using Jensen's inequality, we can easily show that $E_{\theta^\star}\mathcal{L}_\theta(X) \geq 0$, and that the equality holds if and only if $\theta = \theta^\star$. Therefore, the true value $\theta^\star$ minimizes the expected risk. The empirical risk is equal to

$$\widehat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^\star}(Z_i)}{p_\theta(Z_i)}.$$

The numerators $p_{\theta^\star}(Z_i)$ are irrelevant for the minimization of the empirical risk, so we can equivalently maximize the simplified function

$$\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(Z_i).$$

The MLE estimator $\widehat{\theta}_n$ is then defined as any element of the set

$$\arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(Z_i).$$

2. Binary Classification: suppose that the points $Z_1, \ldots, Z_n$ are pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$. This problem is often encountered in practice. For example, the variable X could denote the levels of particular biomarkers in a person's blood, and Y could encode whether that person has a particular disease ($Y = 1$) or not ($Y = 0$). The goal of binary classification is to determine a function $f : \mathcal{X} \to \{0, 1\}$ in some function class $\mathscr{F}$ such that the probability of misclassification, $P(f(X) \neq Y)$ is minimized. In that case, the loss function would be defined as

$$\mathcal{L}_f(x, y) := \mathbb{1}_{\{f(x)\neq y\}} = \begin{cases} 1, & \text{if } f(x) \neq y \\ 0, & \text{else} \end{cases} \tag{2.9}$$

Any function $f^\star$ that minimizes $\mathbb{E}\left[\mathcal{L}_f(X, Y)\right] = P(f(X) \neq Y)$ is known as a *Bayes classifier*. It can be easily shown that, if the two classes are marginally equally likely, i.e. $P(Y = 1) =$

$P(Y=0)$, then the function

$$f^\star(x) = \begin{cases} 1, & \text{if } \frac{P(Y=1|X=x)}{P(Y=0|X=x)} \geq 1 \\ 0, & \text{else.} \end{cases} \tag{2.10}$$

is a Bayes classifier. However, $P(Y=1), P(Y=0), P(Y=1 \mid X=x)$ and $P(Y=0 \mid X=x)$ are unknown, so we do not have access to the Bayes classifier. Therefore, we use the ERM estimator, which is now defined as

$$\widehat{f}_n \in \underset{f\in\mathscr{F}}{\arg\min}\, \widehat{R}_n(f) = \underset{f\in\mathscr{F}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{f(X_i)\neq Y_i\}}\right]. \tag{2.11}$$

3. Regression: suppose that the points $Z_1, \ldots, Z_n$ are again pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are now determined by the equation $Y = \mu_0(X) + \varepsilon$. Here, $\mu_0 : \mathcal{X} \to \mathbb{R}$ is an unknown map that is referred to as the *regression* function, and ε is a zero-mean random variable (noise) that is independent of (X, Y). Once more, we assume that the regression funcion belongs to a function class $\mathscr{F}$. Since $\mathbb{E}[\varepsilon] = 0$, it follows that $\mu_0(x) = \mathbb{E}[Y \mid X = x]$. For any $\mu \in \mathscr{F}$, we define the loss function $\mathcal{L}_\mu(x, y) = (y - \mu(x))^2$. It can be easily shown that the true regression function μ_0 minimizes $\mathbb{E}\left[\mathcal{L}_\mu(X, Y)\right]$ over $\mathscr{F}$. However, μ_0 is unknown, so we estimate it using the ERM estimator

$$\widehat{\mu}_n \in \underset{\mu\in\mathscr{F}}{\arg\min}\, \widehat{R}_n(\mu) = \underset{\mu\in\mathscr{F}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu(X_i))^2\right]. \tag{2.12}$$

This estimator is known as the *least squares estimator*.

By definition, the ERM estimator $\widehat{f}_n$ depends on the data points $Z_1, \ldots, Z_n$. The reason why this is problematic will become clear in the next paragraph.

An interesting question that one could ask is the following: how large is $R\left(\widehat{f}_n, f^\star\right)$, i.e. the (theoretical) risk of the empirical risk minimizer? Ideally, since $R(f, f^\star)$ is the quantity that we actually want to minimize (and not $\widehat{R}_n(f)$), we would like that $\widehat{f}_n$ satisfies

$$R\left(\widehat{f}_n, f^\star\right) \approx \inf_{f\in\mathscr{F}} R(f, f^\star),$$

at least for large values of n. To investigate whether this holds, we can temporarily assume that $\inf_{f\in\mathscr{F}} R(f, f^\star)$ is achieved at some point f_0 - not necessarily equal to $f^\star$. We define the *excess risk* of $\widehat{f}_n$ by

$$E\left(\widehat{f}_n, f^\star\right) := R\left(\widehat{f}_n, f^\star\right) - R(f_0, f^\star). \tag{2.13}$$

It obviously holds that $E\left(\widehat{f}_n, f^\star\right) \geq 0$. We essentially want to show that $E\left(\widehat{f}_n, f^\star\right) \overset{n\to\infty}{\longrightarrow} 0$. Notice that the excess risk can be decomposed as

$$E\left(\widehat{f}_n, f^\star\right) \quad = \quad \left[R\left(\widehat{f}_n, f^\star\right) - \widehat{R}_n\left(\widehat{f}_n\right)\right] + \left[\widehat{R}_n\left(\widehat{f}_n\right) - \widehat{R}_n\left(f_0\right)\right] + \left[\widehat{R}_n\left(f_0\right) - R\left(f_0, f^\star\right)\right]. \quad (2.14)$$

The second term is non positive since $\widehat{f}_n$ minimizes the empirical risk. The third term is equal to

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{f_0}\left(Z_i\right)\right] - \mathbb{E}_{f^\star}\mathcal{L}_{f_0}\left(Z\right),$$

which converges to zero a.s. by the SLLN. A naive first approach would be to use the same argument (SLLN) for the first term, which can be written as

$$\mathbb{E}_{f^\star}\mathcal{L}_{\widehat{f}_n}(Z) - \left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\widehat{f}_n}(Z_i)\right].$$

However, this would not be correct, since $\widehat{f}_n$ is a random quantity that depends on $Z_1, \ldots, Z_n$ and varies with n. Therefore, it is not possible to use the SLLN to prove that this term converges to zero. However, it holds that

$$\left|R\left(\widehat{f}_n, f^\star\right) - \widehat{R}_n\left(\widehat{f}_n\right)\right| \leq \sup_{f\in f}\left|R\left(f, f^\star\right) - \widehat{R}_n\left(f\right)\right|.$$

Therefore, one potential strategy would be to show that

$$\sup_{f\in\mathscr{F}}\left|R\left(f, f^\star\right) - \widehat{R}_n\left(f\right)\right| \overset{n\to\infty}{\longrightarrow} 0, \ P\text{-a.s.} \quad (2.15)$$

Notice that

$$\begin{aligned} R\left(f, f^\star\right) - \widehat{R}_n\left(f\right) &= \mathbb{E}_{f^\star}\mathcal{L}_{\widehat{f}_n}(Z) - \left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\widehat{f}_n}(Z_i)\right] \\ &= \int_{\mathcal{Z}}\mathcal{L}_f\, dP_Z - \int_{\mathcal{Z}}\mathcal{L}_f\, d\mathbb{P}_n \\ &= \int_{\mathcal{Z}}\mathcal{L}_f\, d\left(P_Z - \mathbb{P}_n\right). \end{aligned} \quad (2.16)$$

Motivated by this discussion, we give the following definition.

Definition 2.1.2. Let $Z_1, Z_2, \ldots : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{G})$ be a sequence of i.i.d. random elements and let $\mathscr{F}$ be a class of real-valued measurable functions $\mathcal{Z} \to \mathbb{R}$. We say that $\mathscr{F}$ satisfies the *Uniform Law of*

Large Numbers (ULLN) if it holds that

$$\sup_{f\in\mathscr{F}}\left|\int_{\mathcal{Z}} f\, d(\mathbb{P}_n - P_Z)\right| \xrightarrow{P} 0, \text{ as } n\to\infty \tag{2.17}$$

We shall use the notation

$$\|\mathbb{P}_n - P_Z\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}}\left|\int_{\mathcal{Z}} f\, d(\mathbb{P}_n - P_Z)\right|$$

Remark 2.1.3. Equations (2.16) and (2.17) might cause some confusion since f denotes different things in each of them. This remark aims to clarify the notation and explain this inconsistency.

- For the purposes of ERM, we are interested in the function class $\mathscr{L}_{\mathscr{F}} := \{\mathcal{L}_f : f \in \mathscr{F}\}$, which is parametrized by $\mathscr{F}$. This becomes obvious from Equations (2.14) and (2.16).
- However, we can discuss the ULLN in its full generality, outside of the spectrum of ERM. In that case we can allow for any general function class $\mathscr{F}$.

Therefore, the rule in the rest of this thesis is the following: we discuss the GC function classes in their full generality using the symbol $\mathscr{F}$. When going back to the specific area of ERM, we switch from $\mathscr{F}$ to $\mathscr{L}_{\mathscr{F}}$.

From the above discussion it becomes clear that, the ULLN is a sufficient condition for the convergence of the excess risk to zero (in probability). Therefore, it is necessary to investigate which function classes satisfy the ULLN. The first result in this direction was proved in 1933 independently by Glivenko and Cantelli [Glivenko, 1933, Cantelli, 1933].

Theorem 2.1.4 (Glivenko-Cantelli)**.** Let $Z_1, \dots, Z_n : \Omega \to \mathbb{R}$ be i.i.d. random variables with cumulative distribution function F_Z. Let $\mathbb{F}_n$ be the empirical c.d.f. of $Z_1, \dots, Z_n$. Then, it holds that

$$\|\mathbb{F}_n - F_Z\|_\infty = \sup_{z\in\mathbb{R}} |\mathbb{F}_n(z) - F_Z(z)| \xrightarrow{a.s.} 0. \tag{2.18}$$

Although it might not be obvious how this theorem is related to the ULLN, a closer look reveals that

$$\mathbb{F}_n(z) - F_Z(z) = \int_{\mathbb{R}} \mathbb{1}_{(-\infty,z]}\, d\left(\mathbb{F}_n - F_Z\right).$$

Therefore, this theorem essentially tells us that the function class $\left\{\mathbb{1}_{(-\infty,z]} : z \in \mathbb{R}\right\}$ satisfies the ULLN. Due to this seminal result, the function classes that satisfy the ULLN are alternatively called *Glivenko-Cantell (GC) classes*.

To conclude, the main question that we try to investigate in this chapter is the following: under what conditions is a function class a GC class? As we shall see, this is closely related to the size and the complexity of the class. Therefore, we will focus on developing measures of size and complexity that give insight into the structure of a function class and provide necessary conditions for a class to satisfy the ULLN.

Throughout this chapter, we follow the exposition of [Wainwright, 2019], with several ideas also taken by [van de Geer, 2000] and [Vershynin, 2018]. We emphasize that, although the definition of a GC class only requires convergence in probability, the results we are going to present often yield almost sure convergence, which is stronger.

2.2 Concentration bounds

Before moving forward to the investigation of complexity measures for function classes, it is necessary to take a step back and study the spread of random variables around their means, as well as in the tails of their distribution. Throughout this section, we use the following terminology:

1. **Tail bound:** An upper bound on the probability that a random variable takes values at the tails of its distribution (i.e. very large or very small values). The simplest and most general tail bound is given by Markov's inequality:

   for any nonnegative random variable X and any $a > 0$, it holds that

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

2. **Concentration bound:** An upper bound on the probability that a random variable takes values *away* from its mean. One of the most well-known concentration bounds is Chebyshev's inequality:

   for any random variable X with finite first and second moments, and any $\kappa > 0$, it holds that

   $$P(|X - \mathbb{E}[X]| > \kappa) \leq \frac{\mathrm{Var}(X)}{\kappa^2}.$$

This section is going to provide us with some concentration bounds that will be extensively used in the proofs of the upcoming results. Our goal in the next sections will be to show that

$$\sup_{f \in \mathcal{F}} |R_n(f) - R(f, f^\star)| \approx \mathbb{E}\left[\sup_{f \in \mathcal{F}} |R_n(f) - R(f, f^\star)|\right], \tag{2.19}$$

which is nothing more than a concentration property of the random quantity

$$\sup_{f \in \mathcal{F}} \left| R_n(f) - R(f, f^\star) \right|.$$

The concentration bounds that we prove in this section will turn out to be very useful in showing concentration properties of this form.

The upcoming results are not exclusively used in Empirical Process Theory. On the contrary, they have a very large field of application, ranging from Concentration of Measure, to Optimal Transport, Optimization, and Stochastic Calculus. Therefore, even though this section might initially seem to lie out of the spirit of this thesis, it contains a lot of results of independent interest. The importance of this results in Empirical Process Theory will become very clear in Sections 2.3 and 2.4.

Markov's and Chebyshev's inequality have some interesting extensions. One of them is the so-called *Chernoff bound*, which is discussed in Example 2.2.1.

Example 2.2.1 (Chernoff bound)**.** Let X be any random variable with mean μ. Suppose that there exists a constant b such that the moment generating function $\mathbb{E}\left[e^{\lambda X}\right]$ is finite for all $\lambda \in [0, b]$. From Markov's inequality, it follows that, for all $\lambda \in \mathbb{R}$,

$$\begin{aligned} P\left[X - \mu \geq t\right] &= P\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \\ &\leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}. \end{aligned}$$

This inequality can be rewritten as

$$\log P\left[X - \mu \geq t\right] \leq \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t.$$

Since $\lambda \in [0, b]$ was arbitrary, the latter implies that

$$\log P\left[X - \mu \geq t\right] \leq \inf_{\lambda \in [0,b]} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\}. \tag{2.20}$$

This last inequality is widely known as the Chernoff bound.

Under distributional assumptions, it is often easy to derive further concentration and tail bounds. Example 2.2.2 illustrates the derivation of such bounds for Gaussian random variables.

Example 2.2.2 (Gaussian Concentration Bound)**.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable. It is well known that the moment-generating function of X is given by $\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$.

From the Chernoff bound, it follows that

$$\log P\left[X-\mu \geq t\right] \leq \inf_{\lambda \in [0,\infty)} \left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right\} = \inf_{\lambda \geq 0}\left[\frac{\lambda^2\sigma^2}{2} - \lambda t\right] = -\frac{t^2}{2\sigma^2},$$

where, in the last step, we computed the infimum by setting the derivative $\frac{d}{d\lambda}\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right)$ to zero. It follows that

$$P\left[X-\mu \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}. \tag{2.21}$$

Since $-X \sim \mathcal{N}(-\mu, \sigma^2)$, the same argument yields that $P\left[-X+\mu \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}$. Combining these two bounds yields the concentration bound

$$P\left[|X-\mu| \geq t\right] \leq 2e^{-\frac{t^2}{2\sigma^2}}. \tag{2.22}$$

To obtain concentration properties for more general classes of random variables, we need to impose some assumptions on the tail behaviour of the underlying distribution. This already becomes clear from Chebyshev's inequality, where it is necessary to assume the existence of the first two moments. A large group of concentration results concerns *sub-Gaussian* random variables.

Definition 2.2.3. A random variable X with mean $\mu = \mathbb{E}\left[X\right]$ is sub-Gaussian if there exists a positive constant σ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\lambda^2\sigma^2/2}, \text{ for all } \lambda \in \mathbb{R}. \tag{2.23}$$

The constant σ is referred to as the *sub-Gaussian parameter* of X.

Remark 2.2.4. For Gaussian random variables, the bound in (2.23) holds with equality. Therefore, Gaussian random variables are sub-Gaussian, with their standard deviation as the sub-Gaussian parameter. However, there exist sub-Gaussian random variables that are not Gaussian. For example, it is a standard result that any bounded random variable $X \in [a, b]$ is sub-Gaussian with subgaussian parameter equal to $\sigma = (b-a)/2$.

Remark 2.2.5. Intuitively, a random variable is sub-Gaussian if the tails of its distribution decay at least as fast as the tails of a Gaussian distribution. Indeed, using the definition of a sub-Gaussian random variable, we can compose an argument similar to the one in Example 2.2.2 and show that any sub-Gaussian random variable satisfies the bounds (2.21) and (2.22). The last step of that argument is still vaild because, for any sub-Gaussian random variable X, it holds that $-X$ is also sub-Gaussian with the same parameter. [Vershynin, 2018, Proposition 2.5.2] some interesting properties related to the moments and the tails of sub-Gaussian random variables.

We are now ready to state Hoeffding's inequality, an inequality that plays a central role in studying concentration properties of sub-Gaussian random variables.

Theorem 2.2.6 (Hoeffding Bound)**.** Let $n \geq 1$ be an integer, and let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with means $\mu_1, \ldots, \mu_n$ and sub-Gaussian parameters $\sigma_1, \ldots, \sigma_n$ respectively. Then, for all $t \geq 0$, it holds that

$$P\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}. \tag{2.24}$$

Proof. The proof relies on the fact that linear operations preserve sub-Gaussianity. In particular, we show that $X_1 + \ldots + X_n$ is sub-Gaussian with mean $\mu_1 + \ldots + \mu_n$ and sub-Gaussian parameter $\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$. Since $X_1, \ldots, X_n$ are independent, it follows that

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i)}\right] &= \mathbb{E}\left[\exp\left\{\lambda(\sum_{i=1}^{n}(X_i - \mu_i)\right\}\right] \\
&= \prod_{i=1}^{n}\mathbb{E}\left[exp\left\{\lambda(X_i - \mu_i)\right\}\right] \\
&\leq \prod_{i=1}^{n}\exp\left\{\frac{\lambda^2\sigma_i^2}{2}\right\} \\
&= \exp\left\{\frac{\lambda^2(\sigma_1^2 + \ldots + \sigma_n^2)}{2}\right\},
\end{aligned} \tag{2.25}$$

where we used the fact that $X_1, \ldots, X_n$ are sub-Gaussian in the third step. This shows that $X_1 + \ldots + X_n$ is sub-Gaussian. Inequality (2.24) now follows from the bound (2.21), which, as we discussed in Remark 2.2.5, holds for all sub-Gaussian random variables. □

Hoeffding inequality has an interesting extension, known as the Azuma-Hoeffding inequality. This inequality was introduced by [Azuma, 1967] and it extends the Hoeffding bound to random variables that have a non-trivial dependence structure. More specifically, it covers the case of *martingale difference* sequences. If $\{\mathcal{A}_i\}_{i=1}^{\infty}$ is a filtration in a probability space $(\Omega, \mathcal{A}, P)$, then we say that a sequence $\{D_i\}_{i=1}^{\infty}$ of random variables is a martingale difference sequence (MDS) if, for all $i \geq 1$, it holds that D_i is $\mathcal{A}_i$-measurable, integrable, and satisfies the condition

$$\mathbb{E}\Big[D_i \mid \mathcal{A}_{i-1}\Big] = 0. \tag{2.26}$$

Theorem 2.2.7 (Azuma-Hoeffding)**.** Let $\{D_i\}_{i=1}^{\infty}$ be a martingale difference sequence. If, for all $i \geq 1$, the random variable D_i lies almost surely within an interval of length L_i, then for all $n \geq 1$ it holds that

$$P\left(\sum_{i=1}^{n} D_i \geq t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}\right\} \tag{2.27}$$

Proof. Let $i \geq 1$ be an arbitrary index. Since D_i lies almost surely within an interval of length L_i, the same holds for the conditioned random variable $(D_i \mid \mathcal{A}_{i-1})$. Therefore, from Remark 2.2.4, it follows that $(D_i \mid \mathcal{A}_{i-1})$ is sub-Gaussian with parameter $\sigma = L_i/2$. The mean of this random variable is equal to zero, so, by the definition of a sub-Gaussian random variable, it follows that

$$\mathbb{E}\left[e^{\lambda D_i} \mid \mathcal{A}_{i-1}\right] \leq e^{\lambda^2 L_i^2/8}. \tag{2.28}$$

We now use the law of iterated expectation to derive a decomposition of $\mathbb{E}\left[\exp\left\{\sum_{i=1}^{n} D_i\right\}\right]$ in a similar way as in the proof of Theorem 2.2.6. We have

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n} D_i\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} D_i\right\} \cdot \exp\left\{\lambda D_n\right\}\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} D_i\right\} \cdot \exp\left\{\lambda D_n\right\} \,\middle|\, \mathcal{A}_{n-1}\right]\right) \\
&= \mathbb{E}\left(\exp\left\{\lambda \sum_{i=1}^{n-1} D_i\right\} \cdot \mathbb{E}\left[\exp\left\{\lambda D_n\right\} \mid \mathcal{A}_{n-1}\right]\right) \\
&\leq \mathbb{E}\left(\exp\left\{\lambda \sum_{i=1}^{n-1} D_i\right\} \cdot \exp\left\{\lambda^2 L_n^2/8\right\}\right) \\
&= \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} D_i\right\}\right] \cdot \exp\left\{\lambda^2 L_n^2/8\right\},
\end{aligned} \tag{2.29}$$

where we used the law of iterated expectation in the second step, the fact that $\exp\left\{\sum_{i=1}^{n-1} D_i\right\}$ is $\mathcal{A}_{n-1}$-measurable in the third step, and Equation 2.28 in the fourth step. Iterating this argument over the remaining terms yields that

$$\mathbb{E}\left[\exp\left\{\sum_{i=1}^{n} D_i\right\}\right] \leq \exp\left\{\frac{\lambda^2}{8} \sum_{i=1}^{n} L_i^2\right\}.$$

This shows that the random variable $\sum_{i=1}^{n} D_i$ is sub-Gaussian with parameter $\sigma = \frac{1}{2}\sqrt{\sum_{i=1}^{n} L_i^2}$. From the Remark 2.2.5, and from the bound (2.21), it follows that

$$P\left(\sum_{i=1}^{n} D_i \geq t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}\right\},$$

which finishes the proof. □

Recall that the quantity that we would like to bound is

$$\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P_Z) f| ,$$

which can be viewed as a function $g_n : \mathcal{Z}^n \to \mathbb{R}$. In Theorem 2.2.9, we use the Azuma-Hoeffding bound to show a concentration property for functions of this type that satisfy a certain assumption, called the assumption of *bounded differences*.

Definition 2.2.8. Given two vectors $z, z' \in \mathcal{Z}^n$ and an index $k \in \{1, \dots, n\}$, we define a new vector $z^{\setminus k} \in \mathcal{Z}^n$ by

$$z_j^{\setminus k} := \begin{cases} z_j, & \text{if } j \neq k \\ z'_j, & \text{if } j = k \end{cases}$$

for all $j = 1, \dots, n$. In other words, $z^{\setminus k}$ has all its entries equal to the ones of z except from the k-th one, which is equal to the k-th entry of z'. We say that a function $g : \mathcal{Z}^n \to \mathbb{R}$ has the property of *bounded differences* if, for any index $k \in \{1, \dots, n\}$, there exists a constant $L_k > 0$ such that

$$\left| g(z) - g\left(z^{\setminus k}\right) \right| \leq L_k, \text{ for all } z, z' \in \mathcal{Z}^n. \tag{2.30}$$

Theorem 2.2.9 (McDiarmid Bound)**.** Let $g_n : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function that satisfies the property of bounded differences with constants $L_1, \dots, L_n$. If $Z \in \mathcal{Z}^n$ is a random vector with independent entries, then, for all $t > 0$, it holds that

$$P\Big[g_n(Z) - \mathbb{E} g_n(Z) \geq t \Big] \leq \exp\left\{ -\frac{2t^2}{\sum_{k=1}^n L_k^2} \right\}. \tag{2.31}$$

Proof. Our goal is to decompose the quantity $g_n(Z) - \mathbb{E} g_n(Z)$ into a partial sum of a martingale sequence and then apply the Azuma-Hoeffding bound. Therefore, for any $k \in \{2, \dots, n\}$, we define

$$D_i := \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_i \right] - \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1} \right].$$

We also define $D_1 := \mathbb{E}\left[g_n(Z) \mid Z_1 \right] - \mathbb{E} g_n(Z)$. Obvisously, it holds that $g_n(Z) - \mathbb{E} g_n(Z) = \sum_{i=1}^n D_i$. Also, it is easy to see that $\{D_i\}_{i=1}^n$ forms a martingale difference sequence with respect to the filtration $\{\sigma(Z_1, \dots, Z_i)\}_{i=1}^n$. To apply the Azuma-Hoeffding bound, we still need to show that $D_1, \dots, D_n$ are

bounded. For each $i \in \{1, \dots, n\}$, we define

$$A_i := \inf_{z \in \mathcal{Z}} \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}, z\right] - \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}\right],$$
$$B_i := \sup_{z \in \mathcal{Z}} \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}, z\right] - \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}\right].$$

For all $i \in \{1, \dots, n\}$, it holds that $A_i \leq D_i \leq B_i$. We are now going to show that $B_i - A_i \leq L_i$, where L_i is the constant that appears in the property of bounded differences. Due to the independence of the entries of Z, it follows that [3]

$$\mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_i\right] = G_{n,i}\left(Z_1, \dots, Z_i\right),$$

where $G_{i,n}\left(z_1, \dots, z_i\right) = \mathbb{E}\left[g_n\left(z_1, \dots, z_i, Z_{i+1}, \dots, Z_n\right)\right]$. It follows that

$$\begin{aligned} B_i - A_i &= \sup_{z \in \mathcal{Z}} \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}, z\right] - \inf_{z \in \mathcal{Z}} \mathbb{E}\left[g_n(Z) \mid Z_1, \dots, Z_{i-1}, z\right] \\ &\leq \sup_{z, z' \in \mathcal{Z}} \left| G_{n,i}\left(Z_1, \dots, Z_{i-1}, z\right) - G_{n,i}\left(Z_1, \dots, Z_{i-1}, z'\right) \right|. \end{aligned} \tag{2.32}$$

However, from the property of bounded differences, it directly follows that

$$\sup_{z, z' \in \mathcal{Z}} \left| G_{n,i}\left(z_1, \dots, z_{i-1}, z\right) - G_{n,i}\left(z_1, \dots, z_{i-1}, z'\right) \right| \leq L_i$$

for all $z_1, \dots, z_{i-1} \in \mathcal{Z}$. Given Equation (2.32), the latter observation implies that

$$B_i - A_i \leq L_i,$$

which is exactly what we wanted to prove. The McDiarmid bound now follows directly from the Azuma-Hoeffding inequality. □

2.3 Rademacher Complexity

Whether a function class $\mathscr{F}$ is a GC class naturally depends on the size and the complexity of the class. For example, if $\mathscr{F}$ contains only one element $f : \mathcal{Z} \to \mathbb{R}$, then

$$\|\mathbb{P}_n - P_Z\|_{\mathscr{F}} = \left| \int_{\mathcal{Z}} f \, d\left(\mathbb{P}_n - P_Z\right) \right|,$$

[3] This is a well-known property of the conditional expectation, which can be found in most of the standard probability textbooks, e.g. [Durrett, 2019, Example 4.1.7].

which converges to zero due to the SLLN. Similarly, we can show that finite function classes satisfy the ULLN. Moreover, as the Glivenko-Cantelli theorem shows, there are also infinite classes that satisfy the ULLN. However, there exist function classes that are not GC. Example 2.3.1, borrowed from [Wainwright, 2019], presents such a class.

Example 2.3.1. Let $\mathcal{S}$ be the family of all finite subsets of $\mathbb{R}$ and let $\mathscr{F}_{\mathcal{S}} = \{\mathbb{1}_S : S \in \mathcal{S}\}$ be the family of the indicator functions of these sets. We shall show that $\mathscr{F}_{\mathcal{S}}$ is not a GC class. Suppose that the distribution P_Z of $Z_1, Z_2, \ldots$ has no atoms, i.e. $P_Z(\{z\}) = 0$ for all $z \in \mathbb{R}$. Then, for any $S \in \mathcal{S}$, it holds that $P_Z(S) = 0$, since S is finite. However, for any $n \in \mathbb{N}$, the finite set $S_n := \{Z_1, \ldots, Z_n\}$ belongs to $\mathcal{S}$, and it clearly holds that $\mathbb{P}_n(S_n) = 1$. Therefore,

$$\begin{aligned}
\|\mathbb{P}_n - P_Z\|_{\mathscr{F}_{\mathcal{S}}} &= \sup_{f \in \mathscr{F}_{\mathcal{S}}} |(\mathbb{P}_n - P_Z) f| \\
&= \sup_{S \in \mathcal{S}} |(\mathbb{P}_n - P_Z) \mathbb{1}_S| \\
&\geq |(\mathbb{P}_n - P_Z) \mathbb{1}_{S_n}| \\
&= 1,
\end{aligned}$$

which shows that $\mathscr{F}_{\mathcal{S}}$ does not satisfy the ULLN.

We should interpret the result of the previous example as an indication that the function class $\mathscr{F}_{\mathcal{S}}$ is too large or too complex for the ULLN to hold. How should we measure the size and the complexity of a function class, especially when it has infinitely many elements? Which measures of size and complexity are informative as to whether a particular function class is a GC class? Our task in this section is to develop such measures and to investigate their sufficiency in deriving such properties of function classes. The first of these measures is the *Rademacher complexity*, which is defined in Definition 2.3.2.

Definition 2.3.2. Let $Z_1, \ldots, Z_n : \Omega \to \mathcal{Z}$ be i.i.d. random elements, and let $\mathscr{F}$ be a class of functions $\mathcal{Z} \to \mathbb{R}$. The (empirical) Rademacher complexity of $\mathscr{F}$ with respect to P_Z [4] is defined as

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_{Z,\varepsilon}\left[\sup_{f \in \mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right], \tag{2.33}$$

where $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of Rademacher random variables [5] defined on Ω and independent from $Z_1, \ldots, Z_n$.

[4] The Rademacher complexity depends on the underlying distribution P_Z, but in this thesis we do not focus on this dependence. Instead, we always consider the distribution P_Z as fixed, and we derive the properties of the Rademacher complexity with respect to this distribution.

[5] A Rademacher random variable takes only the values ± 1, each with probability equal to $1/2$.

Notice that $\sum_{i=1}^{n} \varepsilon_i f(X_i)$ is equal to the correlation of the vectors $(\varepsilon_1, \ldots, \varepsilon_n)$ and $(f(Z_1), \ldots, f(Z_n))$. The idea behind this definition is that, if $\mathscr{F}$ is large (or complex) enough, then, for any randomly drawn vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, there should exist a function $f \in \mathscr{F}$ such that $(f(Z_1), \ldots, f(Z_n))$ has a high correlation with ε. Therefore, large and complex function classes should have a large Rademacher complexity. The Rademacher complexity has historically been studied as a complexity measure in the context of Banach spaces [Milman et al., 2001] and lately also in the context of Empirical Risk Minimization [Van Der Vaart and Wellner, 1996]. More references related to the Rademacher complexity can be found in [Wainwright, 2019].

The reason why the Rademacher complexity is so closely related to the ULLN becomes clear from Theorem 2.3.3 [Wainwright, 2019, Theorem 4.10], which we prove below. We say that a function class $\mathscr{F}$ is b-uniformly bounded if, for all elements $f : \mathcal{Z} \to \mathbb{R}$ it holds that $\sup_{z \in \mathcal{Z}} |f(z)| \leq b$.

Theorem 2.3.3. Let $\mathscr{F}$ be a b-uniformly bounded class of functions $\mathcal{Z} \to \mathbb{R}$. Then, for any integer $n \geq 1$ and any real number $\delta > 0$, it holds that

$$\|\mathbb{P}_n - P_Z\|_{\mathscr{F}} \leq 2\mathcal{R}_n(\mathscr{F}) + \delta \tag{2.34}$$

with P_Z-probability at least $1 - \exp\left(-\frac{n\delta^2}{2b^2}\right)$.

Proof. The proof of the theorem consists of two parts. First we show that $\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}$ is tightly concentrated around its mean, and then we derive an upper bound for this mean.

Concentration around mean: To simplify the notation, we consider the recentered functions $\overline{f}(z) := f(z) - \mathbb{E}[f(Z)]$. Then, it holds that $\|\mathbb{P}_n - P\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(Z_i) \right|$. We consider the function $G : \mathcal{Z}^n \to \mathbb{R}$, defined as

$$G(z_1, \ldots, z_n) := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(z_i) \right|.$$

We show that G has the property of bounded differences that was defined in Definition 2.2.8. Since G is symmetric, it only suffices to check that this property is satisfied for the first coordinate. We define a vector $y \in \mathbb{R}^n$ which differs from x only in the first coordinate. Notice that

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(x_i) \right| - \sup_{h \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \overline{h}(y_i) \right| &\leq \left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(y_i) \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{n} \overline{f}(x_i) - \frac{1}{n} \sum_{i=1}^{n} \overline{f}(y_i) \right| \\
&= \frac{1}{n} \left| \overline{f}(x_1) - \overline{f}(y_1) \right|
\end{aligned}$$

$$\leq \frac{2b}{n}, \tag{2.35}$$

where the last inequality uses the uniform boundedness of f and the fact that $\overline{f}(x_1) - \overline{f}(y_1) = f(x_1) - f(y_1)$. Since the inequality (2.35) holds for all $f \in \mathscr{F}$, we can take the supremum over $\mathscr{F}$. This yields that

$$\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \overline{f}(x_i) \right| - \sup_{h\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \overline{h}(y_i) \right| \leq \frac{2b}{n},$$

which is equivalent to $G(x) - G(y) \leq 2b/n$. Interchanging x, y and using the same argument yields that $G(y) - G(x) \leq 2b/n$, so we deduce that $|G(x) - G(y)| \leq 2b/n$. From Lemma 2.2.9 (McDiarmid bound), it follows that, for all $t \geq 0$,

$$G(X_1, \ldots, X_n) - \mathbb{E}G(X_1, \ldots, X_n) \leq t$$

with P-probability at least $1 - \exp\left(-\frac{nt^2}{2b^2}\right)$. From the definition of G, this is equivalent to

$$\|\mathbb{P}_n - P_Z\|_{\mathscr{F}} - \mathbb{E}\left(\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}\right) \leq t$$

with probability at least $1 - \exp\left(-\frac{nt^2}{2b^2}\right)$.

Upper bound on $\mathbb{E}\left(\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}\right)$: We now show that the mean is upper bounded by $2\mathcal{R}_n(\mathscr{F})$. We use *symmetrization*. We consider random variables $Y_1, \ldots, Y_n$ that are independent with each other, independent from $X_1, \ldots, X_n$ and have the same distribution as $X_1, \ldots, X_n$. Then, it holds that

$$\begin{aligned}
\mathbb{E}\left(\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}\right) &= \mathbb{E}\left(\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i) \right|\right) \\
&= \mathbb{E}_X\left(\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(Y_i) \right|\right) \\
&= \mathbb{E}_X\left(\sup_{f\in\mathscr{F}} \left| \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^{n} \{f(X_i) - f(Y_i)\}\right] \right|\right) \\
&\leq \mathbb{E}_{X,Y}\left(\sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \{f(X_i) - f(Y_i)\} \right|\right),
\end{aligned} \tag{2.36}$$

where the last step follows from Jensen's inequality and the convexity of the supremum function[6]. We now want to compare the latter quantity with the Rademacher complexity of $\mathscr{F}$. Therefore, we consider i.i.d. Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$, which are also independent from $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The crucial observation is that, since X_i, Y_i are i.i.d. and independent from ε_i, the

[6] It holds that $\sup_{x\in\mathcal{X}} [\lambda f(x) + (1-\lambda)g(x)] \leq \lambda \sup_{x\in\mathcal{X}} f(x) + (1-\lambda)\sup_{x\in\mathcal{X}} g(x)$.

random vector with entries $\varepsilon_i \left(f(X_i) - f(Y_i)\right)$ has the same distribution as the random vector with entries $f(X_i) - f(Y_i)$ [7]. It follows that

$$\begin{aligned}
\mathbb{E}_{X,Y}\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-f(Y_i)\right\}\right|\right) &= \mathbb{E}_{X,Y,\varepsilon}\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-f(Y_i)\right\}\right|\right)\\
&\leq \mathbb{E}_{X,Y,\varepsilon}\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\\
&= 2\mathbb{E}_{X,\varepsilon}\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\\
&= 2\mathcal{R}_n\left(\mathscr{F}\right).
\end{aligned}$$

□

From this theorem we can deduce that, if $\mathcal{R}_n\left(\mathscr{F}\right) = o(1)$, then $\left\|\mathbb{P}_n - P\right\|_{\mathscr{F}} \xrightarrow{\text{a.s.}} 0$. To prove this, we can use the Borel-Cantelli lemma. Consider the sets

$$A_n := \left\{\left\|\mathbb{P}_n - P_Z\right\|_{\mathscr{F}} > 2\mathcal{R}_n\left(\mathscr{F}\right) + \sqrt{2}b\left(\frac{2\log n}{n}\right)^{1/2}\right\}.$$

From Theorem 2.3.3 it follows that $P\left(A_n\right) \leq \exp\left(-2\log n\right) = 1/n^2$. Since $\sum_{n=1}^{\infty} P\left(A_n\right) = \sum_{n=1}^{\infty} 1/n^2$, it follows from the lemma of Borel-Cantelli that $P\left(\limsup_{n\in\mathbb{N}} A_n\right) = 0$. Consider an element $\omega \in \Omega\backslash\left(\limsup_{n\in\mathbb{N}} A_n\right)$. From the definition of $\limsup_{n\in\mathbb{N}} A_n$, it follows that there exists an integer N_ω such that $\omega \notin A_n$ for all $n \geq N_\omega$. This means that

$$\left\|\mathbb{P}_n(\omega) - P_Z\right\|_{\mathscr{F}} \leq 2\mathcal{R}_n\left(\mathscr{F}\right) + \sqrt{2}b\left(\frac{2\log n}{n}\right)^{1/2}$$

for all $n \geq N_\omega$. Since $\mathcal{R}_n\left(\mathscr{F}\right) = o(1)$, it follows that

$$\left\|\mathbb{P}_n(\omega) - P_Z\right\|_{\mathscr{F}} \xrightarrow{n\to\infty} 0.$$

Since $P\left(\Omega\backslash\limsup_{n\in\mathbb{N}} A_n\right) = 1$, we deduce that $\left\|\mathbb{P}_n - P_Z\right\|_{\mathscr{F}} \xrightarrow{\text{a.s.}} 0$. Therefore, if $\mathcal{R}_n\left(\mathscr{F}\right) = o(1)$, then the class $\mathscr{F}$ is a GC class.

[7] Let $Z \sim F_Z$ be a symmetric random variable and ε a Rademacher random variable, independent from Z. It holds that $P\left(\varepsilon Z \leq z\right) = P\left(\varepsilon = 1\right)P\left(\varepsilon Z \leq z \mid \varepsilon = 1\right) + P\left(\varepsilon = -1\right)P\left(\varepsilon Z \leq z \mid \varepsilon = -1\right) = \frac{1}{2}P\left(Z \leq z\right) + \frac{1}{2}P\left(-Z \leq z\right) = P(Z \leq z)$, where in the last step we used the symmetry of Z.

2.4 Entropy with Bracketing

Entropy with bracketing is a measure of complexity that can yield the ULLN in a very straightforward way. It has a different flavour from the Rademacher complexity and its definition resembles the definition of an open cover of subset of a metric space. Just like the existence of finite open covers guarantees compactness, a finite entropy with bracketing guarantees that a function class is GC.

Let Q be a measure on $(\mathcal{Z}, \mathcal{A})$ and let $p \in [1, \infty)$ be a constant. Recall the definition of the space $L_p(Q)$:

$$L_p(Q) := \left\{ f : \mathcal{Z} \to R \;\middle|\; \int |f|^p \, dQ < \infty \right\}.$$

For any function $f \in L_p(Q)$, we define its $L_p(Q)$ norm as $\|f\|_{p,Q} := \left(\int_{\mathcal{Z}} |f|^p \, dQ\right)^{1/p}$ [8].

Definition 2.4.1 (Entropy with bracketing)**.** Let $\mathscr{F} \subset L_p(Q)$ be a function class and let $\delta > 0$ be a real number. Let $N_{p,B}(\delta, \mathscr{F}, Q)$ be a smallest value of $N \in \mathbb{N}$ such that there exist pairs of functions $\left\{ \left[f_j^L, f_j^U \right] \right\}_{j=1}^N$ with the following properties:

- for all $j = 1, \ldots, N$ it holds that $f_j^L, f_j^U \in L_p(Q)$.
- for all $j = 1, \ldots, N$ it holds that $\left\| f_j^U - f_j^L \right\|_{p,Q} \leq \delta$.
- for any $f \in \mathscr{F}$, there exists an index $j = j(f) \in \{1, \ldots, N\}$ such that

$$f_j^L(z) \leq f(z) \leq f_j^U(z) \text{ for all } z \in \mathcal{Z}.$$

If no such finite collection exists, we define $N_{p,B}(\delta, \mathscr{F}, Q) = \infty$. The δ-entropy with bracketing of $\mathscr{F}$ is then defined as

$$H_{p,B}(\delta, \mathscr{F}, Q) = \log N(\delta, \mathscr{F}, Q). \tag{2.37}$$

The pairs $\left[f_j^L, f_j^U \right]$ are called *brackets* and they are denoted by these square brackets because of the third condition in the above definition.

The entropy with bracketing can also be defined for $p = \infty$. In analysis, it is common practice to use the *essential supremum* of a function to define its supremum norm. However, for the next definition, we define the supremum norm of a function $f : \mathcal{Z} \to \mathbb{R}$ as

$$|f|_\infty := \sup_{z \in \mathcal{Z}} |f(z)|. \tag{2.38}$$

[8] In fact, $\|\cdot\|_{p,Q}$ is a seminorm.

With denote this norm with single instead of double brackets to distinguish it from the standard supremum norm that is defined via the essential supremum. Notice that, unlike $\|\cdot\|_\infty$, the norm $|\cdot|_\infty$ does not depend on the measure Q.

Definition 2.4.2 (Entropy with bracketing for the supremum norm)**.** Let $N_\infty(\delta, \mathscr{F})$ be the smallest value of $N \in \mathbb{N}$ for which there exists a finite collection $\{f_j\}_{j=1}^N \subset \mathscr{F}$ with the property that

$$\sup_{f\in\mathscr{F}} \min_{j=1,\dots,N} |f - f_j|_\infty \leq \delta.$$

If no such collection exists, we define $N_\infty(\delta, \mathscr{F}) = \infty$. The δ-entropy of $\mathscr{F}$ with respect to $|\cdot|_\infty$ is then defined as $H_\infty(\delta, \mathscr{F}) = \log N_\infty(\delta, \mathscr{F})$.

As the next result shows, $H_\infty(\delta, \mathscr{F})$ upper bounds $H_{p,B}(\delta, \mathscr{F}, Q)$ for any $p \geq 1$ and any probability measure Q.

Lemma 2.4.3. Let Q be a probability measure on $\mathcal{Z}$ and let $\mathscr{F} \subset L_p(Q)$ be a function class. Then, for all $\delta > 0$,

$$H_{p,B}(\delta, \mathscr{F}, Q) \leq H_\infty\left(\frac{\delta}{2}, \mathscr{F}\right).$$

Proof. Fix a positive real number δ. If $H_\infty(\delta/2, \mathscr{F}) = \infty$, then there is nothing left to prove. Suppose that $N_\infty(\delta/2, \mathscr{F}) = N < \infty$ and let $\{f_j\}_{j=1}^N \subset \mathscr{F}$ be a collection of functions that satisfies Definition 2.4.2 for the threshold value $\delta/2$. From that definition it follows that, for all $f \in \mathscr{F}$, there exists an index $j \in \{1, \dots, N\}$ such that $|f - f_j|_\infty \leq \delta/2$. This is equivalent to

$$f_j(z) - \frac{\delta}{2} \leq f(z) \leq f_j(z) + \frac{\delta}{2}, \text{ for all } z \in \mathcal{Z}. \tag{2.39}$$

We consider the brackets $\{[f_j - \delta/2, f_j + \delta/2]\}_{j=1}^N$. From Equation (2.39), it follows that these brackets cover the whole function class $\mathscr{F}$ in the sense of Definition 2.4.1. Since $\{f_j\}_{j=1}^N \subset \mathscr{F} \subset L_p(Q)$, it easily follows that the functions $f_j \pm \delta/2$, $j = 1, \dots, N$ also belong to $L_p(Q)$. Finally, it is obvious that, for all $j = 1, \dots, N$,

$$\left\| \left(f_j + \frac{\delta}{2}\right) - \left(f_j - \frac{\delta}{2}\right) \right\|_{p,Q} = \delta.$$

We conclude that $\{[f_j - \delta/2, f_j + \delta/2]\}_{j=1}^N$ is a finite bracketing class for $\mathscr{F}$, which proves that

$$N_{p,B}(\delta, \mathscr{F}, Q) \leq N = N_\infty(\delta/2, \mathscr{F}).$$

□

The importance of entropy with bracketing is demonstrated in the following lemma:

Lemma 2.4.4. Let $Z_1, Z_2, \ldots : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{G})$ be a sequence of i.i.d. random variables with distribution P_Z. If $\mathscr{F} \subset L_1(P_Z)$ and

$$H_{1,B}(\delta, \mathscr{F}, P_Z) < \infty \text{ for all } \delta > 0,$$

the $\mathscr{F}$ satisfies the ULLN.

Proof. Let $\delta > 0$ be an arbitrary real number and let $\left\{\left[f_j^L, f_j^U\right]\right\}_{j=1}^N$ be a bracketing set for $\mathscr{F}$, i.e. a collection that satisfies the assumptions of Definition 2.4.1. Then, for any $f \in \mathscr{F}$, there exists an index $j \in \{1, \ldots, N\}$ such that $f_j^L \leq f \leq f_j^U$. For this index j it holds that

$$\begin{aligned}
\int_{\mathcal{Z}} f\, d(\mathbb{P}_n - P_Z) &= \int_{\mathcal{Z}} f\, d\mathbb{P}_n - \int_{\mathcal{Z}} f\, dP_Z \\
&\leq \int_{\mathcal{Z}} f_j^U\, d\mathbb{P}_n - \int_{\mathcal{Z}} f\, dP_Z \\
&= \int_{\mathcal{Z}} f_j^U\, d(\mathbb{P}_n - P_Z) + \int_{\mathcal{Z}} \left(f_j^U - f\right) dP_Z \\
&\leq \int_{\mathcal{Z}} f_j^U\, d(\mathbb{P}_n - P_Z) + \delta, \qquad (2.40)
\end{aligned}$$

where the last step follows from the fact that $\left\|f_j^U - f_j^L\right\|_{1,P_Z} \leq \delta$ and $f_j^L \leq f \leq f_j^U$. Using a similar argument, we can show that

$$\int_{\mathcal{Z}} f\, d(\mathbb{P}_n - P_Z) \geq \int_{\mathcal{Z}} f_j^L\, d(\mathbb{P}_n - P_Z) - \delta. \qquad (2.41)$$

Since the sets $\left\{f_j^L\right\}_{j=1}^N$ and $\left\{f_j^U\right\}_{j=1}^N$ are finite, it follows directly from the SLLN that there exists an integer $N_0 \geq 1$ such that

$$\max_{j=1,\ldots,N} \left|\int_{\mathcal{Z}} f_j^U\, d(\mathbb{P}_n - P_Z)\right| \leq \delta \quad \text{and} \quad \max_{j=1,\ldots,N} \left|\int_{\mathcal{Z}} f_j^L\, d(\mathbb{P}_n - P_Z)\right| \leq \delta$$

Q-almost surely for all $n \geq N_0$. It follows from Equations (2.40),(2.41) that

$$\sup_{f \in \mathscr{F}} \left|\int_{\mathcal{Z}} f\, d(\mathbb{P}_n - P_Z)\right| \leq 2\delta$$

Q-almost surely. Since δ was arbitrary, we deduce that $\mathscr{F}$ is a GC class. □

Remark 2.4.5. If $p \in [1, \infty)$, then the condition

$$H_{p,B}(\delta, \mathscr{F}, P_Z) < \infty \text{ for all } \delta > 0$$

also yields that the class $\mathscr{F}$ is GC. Indeed, it is a well known fact that, in probability spaces, the L_p norm is increasing in p. Therefore, the condition $H_{p,B}(\delta, \mathscr{F}, P_Z) < \infty$ implies that $H_{1,B}(\delta, \mathscr{F}, P_Z) < \infty$, which yields that the class $\mathscr{F}$ is GC.

Let us now look at some applications of Lemma 2.4.4.

Example 2.4.6 (Classical GC Theorem)**.** As we explained in the discussion after Theorem 2.1.4, the function class $\left\{\mathbb{1}_{(-\infty,z)} : z \in \mathbb{R}\right\}$ satisfies the ULLN. We now show that this is closely relate to Lemma 2.4.4. Let $Z_1, \ldots, Z_n : (\Omega, \mathcal{A}, Q) \to \mathbb{R}$ be i.i.d. random variables with cumulative distribution function F_Z. For simplicity, we assume that F_Z is continuous [9]. Let $\delta > 0$ be an arbitrary real number. Since F_Z is increasing and takes values in the interval $[0, 1]$, there exist points

$$-\infty = a_0 < a_1 < \ldots < a_{n-1} < a_n = +\infty,$$

depending on δ, such that $F_Z(a_{i+1}) - F_Z(a_i) < \delta$ for all $i = 0, \ldots, n-1$. We consider the collection of indicator functions

$$\mathscr{F}_{\text{ind}} = \left\{\left[\mathbb{1}_{(-\infty,a_i]}, \mathbb{1}_{(-\infty,a_{i+1}]}\right]\right\}_{i=0}^{n-1}.$$

Then, the two following conditions hold:

- for all $i = 0, \ldots, n-1$,

$$\begin{aligned}
\left\|\mathbb{1}_{(-\infty,a_i]} - \mathbb{1}_{(-\infty,a_{i+1}]}\right\|_1 &= \int_{\mathbb{R}} \left|\mathbb{1}_{(-\infty,a_i]} - \mathbb{1}_{(-\infty,a_{i+1}]}\right| dP_Z \\
&= \int_{\mathbb{R}} \mathbb{1}_{(-\infty,a_{i+1}]} - \mathbb{1}_{(-\infty,a_i]}\, dP_Z \\
&= \int_{\mathbb{R}} \mathbb{1}_{(-\infty,a_{i+1}]}\, dP_Z - \int_{\mathbb{R}} \mathbb{1}_{(-\infty,a_i]}\, dP_Z \\
&= F_Z(a_{i+1}) - F_Z(a_i) \\
&\leq \delta.
\end{aligned}$$

- for all $z \in \mathbb{R}$, there exists an index $i \in \{0, \ldots, n-1\}$ such that $a_i \leq x < a_{i+1}$. This implies that

$$\mathbb{1}_{(-\infty,a_i]} \leq \mathbb{1}_{(-\infty,x]} \leq \mathbb{1}_{(-\infty,a_{i+1}]}$$

[9] The proof also works similarly in the general case, under some technical tweaks.

pointwise.

These two conditions yield that $H_{1,B}(\delta, \mathscr{F}_{\text{ind}}, P) < \infty$. Since δ was arbitrary, it follows from Lemma 2.4.4 that $\mathscr{F}_{\text{ind}}$ is a GC class.

The above argument is in essence identical to the original proof of the Glivenko-Cantelli theorem. This shows that the use of entropy with bracketing to derive the ULLN is a natural extension of this classical result.

Example 2.4.7 (Bounded Lipschitz Functions)**.** Consider the function class

$$\mathscr{F} := \{f : [0,1] \to [0,1] \mid f \text{ is Lipschitz continuous with Lipschitz constant equal to } L = 1\}.$$

We can show that there exists a constant $A > 0$ such that

$$H_\infty(\delta, \mathscr{F}) \leq \frac{A}{\delta}, \text{ for all } \delta > 0. \tag{2.42}$$

According to Lemma 2.4.3, this implies that $\mathscr{F}$ has a finite bracketing entropy $H_{1,B}(\delta, \mathscr{F}, P_Z)$ with respect to any probability distribution P_Z on $[0,1]$ such that $\mathscr{F} \subset L_1(P_Z)$. In turn, Lemma 2.4.4 implies that $\mathscr{F}$ is a GC class with respect to P_Z [10]. To prove Equation (2.42), we fix a threshold value $\delta > 0$ and we choose a finite sequence

$$0 = a_0 < \ldots < a_N = 1$$

such that $a_j = j\delta$ for all $j \in \{0, \ldots, N-1\}$. We assume that N takes its largest possible value, i.e. $1 - a_{N-1} \leq \delta$. Then, it is easy to see that $N \leq 1 + \lfloor 1/\delta \rfloor$. Denote $B_1 = [a_0, a_1]$ and $B_j = (a_{j-1}, a_j]$ for all $j \in \{2, \ldots, N\}$. Consider a function $f \in \mathscr{F}$. The function $\widetilde{f}$ defined by

$$\widetilde{f} = \sum_{j=1}^{N} \delta \left\lfloor \frac{f(a_j)}{\delta} \right\rfloor \mathbb{1}_{B_j}$$

is piecewise constant and takes values in the set $\Delta := \{0, \delta, \ldots, (N-1)\delta\}$ [11]. It also holds that $\left| f - \widetilde{f} \right|_\infty \leq 2\delta$: indeed, consider a point $z \in [0,1]$ and assume that $z \in B_j$ for some index $j \in$

[10] Notice that, although it makes sense to define the bracketing entropy for any probability measure Q, the ULLN can only be expressed for an i.i.d. sequence of random variables. This is why we referred to a GC class *with respect to the distribution* P_Z of the elements of this sequence. Besides, the condition of Lemma 2.4.4 is also given in this form.

[11] Notice that f takes values in $[0,1]$ so $0 \leq f\left(a_j\right)/\delta \leq 1 < N\delta$. This is why the set of values of $\widetilde{f}$ is upper bounded by $(N-1)\delta$.

$\{1, \ldots, N\}$. Then, from the Lipschitz continuity of f, it follows that

$$\begin{aligned} f(z) - \widetilde{f}(z) &= f(z) - \delta \left\lfloor \frac{f\left(a_j\right)}{\delta} \right\rfloor \\ &\geq f(z) - \delta \cdot \frac{f(a_j)}{\delta} \\ &= f(z) - f(a_j) \\ &\geq -(a_j - z) \\ &\geq -\delta \end{aligned}$$

and

$$\begin{aligned} f(z) - \widetilde{f}(z) &= f(z) - \delta \left\lfloor \frac{f\left(a_j\right)}{\delta} \right\rfloor \\ &\leq f(z) - \delta \left(\frac{f(a_j)}{\delta} - 1 \right) \\ &= f(z) - f(a_j) + \delta \\ &\leq \left(a_j - z\right) + \delta \\ &\leq 2\delta. \end{aligned}$$

It remains to count the number of distince functions $\widetilde{f}$ that can be obtained in this way. This is a pure counting argument that makes some use of the Lipschitz continuity. At first, notice that $\widetilde{f}$ is fully determined by its values at $a_1, \ldots, a_N$.

- $\widetilde{f}\left(a_1\right)$ can take all possible values in the set $\Delta := \{0, \delta, \ldots, (N-1)\delta\}$, so it can take at most $1 + \lfloor 1/\delta \rfloor$ values.
- for any index $j \in \{1, \ldots, N-1\}$, it holds that

$$\begin{aligned} \left|\widetilde{f}(a_{j+1}) - \widetilde{f}(a_j)\right| &\leq \left|\widetilde{f}(a_{j+1}) - f(a_{j+1})\right| + \left|f(a_{j+1}) - f(a_j)\right| + \left|f(a_j) - \widetilde{f}(a_j)\right| \\ &\leq 3\delta. \end{aligned}$$

  Therefore, given $\widetilde{f}\left(a_j\right)$, there are only seven possible values that $\widetilde{f}\left(a_{j+1}\right)$ can take.

From the multiplicative principle, it follows that there exists $(1 + \lfloor 1/\delta \rfloor) \cdot 7^{\lfloor 1/\delta \rfloor}$ distinct choices for $\widetilde{f}$. As we showed, the collection of all these distinct functions satisfies the conditions of Definition

2.4.2 with threshold value equal to 2δ. Replacing δ with $\delta/2$ gives

$$H_\infty(\delta, \mathscr{F}) \leq \log\left(1 + \left\lfloor \frac{2}{\delta} \right\rfloor\right) + \left(1 + \left\lfloor \frac{2}{\delta} \right\rfloor\right) \cdot \log 7,$$

which implies that $H_\infty(\delta, \mathscr{F}) \leq A/\delta$ for a suitable constant $A > 0$.

The last example shows us that sometimes, we might not only be able to show that the bracketing entropy is finite, but even determine specific upper bounds for it. [Birman and Solomjak, 1967] derive such bounds for a variety of function classes, with a specific focus on Sobolev spaces. As we shall see in Section 2.7, bracketing entropy bounds can yield rates of convergence in the ULLN.

2.5 Symmetrization

As we saw in the previous section, the bracketing entropy condition in Lemma 2.4.4 yields uniform laws of large numbers for a variety of function classes. The bracketing entropy is often intractable, so it is usually easier to bound the supremum bracketing entropy $H_\infty(\delta, \mathscr{F})$ and apply Lemma 2.4.3. However, the condition $H_\infty(\delta, \mathscr{F}, P_Z) < \infty$ might sometimes be too restrictive. For instance, as we can easily deduce from Example 2.4.6, the bracketing entropy of the function class $\mathscr{F}_{\text{ind}} := \left\{\mathbb{1}_{(-\infty,z]} : z \in \mathbb{R}\right\}$ is upper bounded by $\log(1 + 1/\delta)$. At the same time, $H_\infty(\delta, \mathscr{F}_{\text{ind}}) = \infty$ for all $\delta < 1$ because, for any two distinct functions in this class, it holds that

$$\sup_{z \in \mathbb{R}} \left| \mathbb{1}_{(-\infty,z_1]} - \mathbb{1}_{(-\infty,z_2]} \right|_\infty = 1.$$

In this section we prove that the ULLN can also be derived from weaker conditions. In particular, we introduce a new notion of entropy, the so called *metric entropy*.

Definition 2.5.1 (Metric entropy)**.** Let (S, d) be a metric space and let $\delta > 0$ an arbitrary positive real number. The *covering number* of S, denoted by $N(\delta, S, d)$, is the minimal number of balls of radius δ that are needed to cover the entire set S. More formally,

$$N(\delta, S, d) = \min\left\{ n \in \mathbb{N} \,\middle|\, \exists \{s_i\}_{i=1}^n \subset S : \sup_{s \in S} \min_{i=1,\dots,n} d(s, s_i) \leq \delta \right\}. \tag{2.43}$$

If no such finite collection of points exists, we set $N(\delta, S, d) = \infty$. We define the *metric entropy* (or simply entropy) by $H(\delta, S, d) = \log N(\delta, S, d)$.

If $S = \mathscr{F} \subset L_p(Q)$ and $d = \|\cdot\|_p$, then we denote the entropy by $H_p(\delta, \mathscr{F}, Q)$. We call $\{s_i\}_{i=1}^n$ a δ-covering set. It is actually not necessary to assume that the elements $s_1, \dots, s_n$ of this set lie in S. More

specifically, it is easy to show that any δ-covering with centers outside of S induces a 2δ-covering with the centers being in S.

The covering number often appears in the literature along the *packing number*.

Definition 2.5.2 (Packing number)**.** Let (S, d) be a metric space and let $\delta > 0$ be an arbitrary positive real number. A δ-packing set is a set $\{s_i\}_{i=1}^{M} \subset S$ such that $d(s_i, s_j) > \delta$ for all distinct indices $i, j \in \{1, \ldots, M\}$. The δ-packing number is the cardinality of the largest δ-packing set and is denoted by $M(\delta, S, d)$.

The covering and the packing numbers are very closely related. This becomes clear from the following lemma.

Lemma 2.5.3. For any $\delta > 0$, it holds that

$$M(2\delta, S, d) \leq N(\delta, S, d) \leq M(\delta, S, d). \tag{2.44}$$

Proof. For the left inequality, we may assume that $N(\delta, S, d) < \infty$, otherwise there is nothing to show. If $\{s_i\}_{i=1}^{N}$ is a δ-covering set with $N = N(\delta, S, d)$, then each ball with center s_i and radius δ can contain at most one point of a (2δ)-packing set. Therefore, the cardinality of any (2δ)-packing set is at most N. This proves the left part of the inequality 2.44.

To show the right part, we may assume that $M(\delta, S, d) < \infty$. If $\{s_i\}_{i=1}^{M}$ is a maximal δ-packing set, then any new point $s \in S \setminus \{s_1, \ldots, s_M\}$ must belong to one of the balls $B(s_1, \delta), \ldots, B(s_M, \delta)$, otherwise it would be possible to extend the δ-packing set. This shows that $\{s_i\}_{i=1}^{M}$, which yields that $N(\delta, S, d) \leq M(\delta, S, d)$. □

In Section 2.4 we worked with L_p spaces, for $p \in [1, \infty)$. In this section, we will make use of the empirical counterpart of the L_p norm. This counterpart is a (random) seminorm, which we denote by $\|\cdot\|_{n,p}$, and which is defined by

$$\|f\|_{n,p} := \left(\frac{1}{n} \sum_{i=1}^{n} |f^p(X_i)| \right)^{1/p}. \tag{2.45}$$

Like before, we denote the entropy of this space by $H_p(\delta, \mathscr{F}, P_n)$. We emphasize that this is a random quantity that depends solely on the random elements $Z_1, \ldots, Z_n$.

The results in this section rely heavily on the technique of *symmetrization*, which is very important for Empirical Process Theory. We could argue that this section is a demonstration of how powerful a simple idea like symmetrization can be. Symmetrization has already been used in the proof of

Theorem 2.3.3, but in this section we are going to formalize it and show its relation to the ULLN more generally.

We consider two i.i.d. random vectors $\mathbb{Z} := (Z_1, \ldots, Z_n)$ and $\mathbb{Z}' = (Z'_1, \ldots, Z'_n)$. The entries of these vectors are also assumed to be i.i.d. with distribution P_Z We denote by $\mathbb{P}_n, \mathbb{P}'_n$ the empirical measures induced by $\mathbb{Z}, \mathbb{Z}'$ respectively. To arrive to our main result, we first have to prove three technical lemmas.

Lemma 2.5.4. Let $\mathbb{P}_n, \mathbb{P}'_n$ and P_Z be as above. Then, it holds that

$$\mathbb{E} \left\| \mathbb{P}_n - P_Z \right\|_{\mathscr{F}} \leq \mathbb{E} \left\| \mathbb{P}_n - \mathbb{P}'_n \right\|_{\mathscr{F}}. \tag{2.46}$$

Proof. Fix a function $f \in \mathscr{F}$. From the independence of $\mathbb{Z}, \mathbb{Z}'$, it follows that $\mathbb{E}\left[\mathbb{P}'_n f \mid \mathbb{Z}\right] = P_Z f$. Also, it is obvious that $\mathbb{E}\left[P_n f \mid \mathbb{Z}\right] = P_n f$. Therefore,

$$\mathbb{E}\left[\left(\mathbb{P}_n - \mathbb{P}'_n\right) f \mid \mathbb{Z}\right] = \left(\mathbb{P}_n - P_Z\right) f.$$

It follows that

$$\begin{aligned}
\left\| \mathbb{P}_n - P_Z \right\|_{\mathscr{F}} &= \sup_{f \in \mathscr{F}} \left| \left(\mathbb{P}_n - P_Z\right) f \right| \\
&= \sup_{f \in \mathscr{F}} \left| \mathbb{E}\left[\left(\mathbb{P}_n - \mathbb{P}'_n\right) f \mid \mathbb{Z}\right] \right| \\
&\leq \mathbb{E}\left(\sup_{f \in \mathscr{F}} \left| \left(\mathbb{P}_n - \mathbb{P}'_n\right) f \right| \,\middle|\, \mathbb{Z} \right) \\
&= \mathbb{E}\left(\left\| \mathbb{P}_n - \mathbb{P}'_n \right\|_{\mathscr{F}} \mid \mathbb{Z} \right),
\end{aligned}$$

where we used Jensen's inequality in the third step. Taking expectations in both sides and using the law of iterated expectations yields that

$$\mathbb{E} \left\| \mathbb{P}_n - P_Z \right\|_{\mathscr{F}} \leq \mathbb{E} \left\| \mathbb{P}_n - \mathbb{P}'_n \right\|_{\mathscr{F}},$$

which finishes the proof. □

Given a function $f \in \mathscr{F}$ and a vector $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$ of i.i.d. Rademacher random variables, independent from $\mathbb{Z}, \mathbb{Z}'$, we define

$$\mathbb{P}_n^{\varepsilon} f = \tfrac{1}{n} \textstyle\sum_{i=1}^n \varepsilon_i f(Z_i) \quad \text{and} \quad \mathbb{P}_n'^{,\varepsilon} f = \tfrac{1}{n} \textstyle\sum_{i=1}^n \varepsilon_i f(Z'_i).$$

The following lemma follows directly from Lemma 2.5.4 and from the fact that, for any $f \in \mathscr{F}$, the vectors

$$\{f(Z_i) - f(Z_i')\}_{i=1}^n \quad \text{and} \quad \{\varepsilon_i (f(Z_i) - f(Z_i'))\}_{i=1}^n$$

have the same distribution. Therefore, its proof is omitted [12].

Lemma 2.5.5. Using the same notation as above, it holds that

$$\mathbb{E} \left\| \mathbb{P}_n - \mathbb{P}_n' \right\|_{\mathscr{F}} \leq 2\mathbb{E} \left\| \mathbb{P}_n^{\varepsilon} \right\|_{\mathscr{F}}. \tag{2.47}$$

Intuitively, the reason why the quantity $\|\mathbb{P}_n^{\varepsilon}\|_{\mathscr{F}}$ is easier to deal with, is that we can control its conditional expectation (given $\mathbb{Z}$) using Hoeffding's inequality. This will become clearer in the proof of the main result of this Section, Theorem 2.5.7.

The first two lemmas provide bounds for the expectation of $\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}$ through the symmetric versions $\|\mathbb{P}_n - \mathbb{P}_n'\|_{\mathscr{F}}$ and $\|\mathbb{P}_n^{\varepsilon}\|_{\mathscr{F}}$. The third lemma complements these results by providing tail bounds for $\|\mathbb{P}_n - P_Z\|_{\mathscr{F}}$. Before moving on to the proof of this lemma, we point out that, for all $f \in \mathscr{F}$ and all $\delta > 0$, it follows from Chebyshev's inequality that

$$P\left(|(\mathbb{P}_n - P_Z) f| > \frac{\delta}{2} \right) \leq \frac{4\mathrm{Var}(f(Z))}{n\delta^2}, \tag{2.48}$$

where the factor n shows up in the denominator due to the fact that

$$\mathrm{Var}\left(\mathbb{P}_n f\right) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n f(Z_i) \right) = \frac{\mathrm{Var}\left(f(Z_1)\right)}{n}.$$

This observation will be useful in the proof of the last technical lemma.

Lemma 2.5.6. Fix $\delta > 0$, and suppose that $n \geq \sup_{f \in \mathscr{F}} \left(8\mathrm{Var}\left(f(Z)\right) / \delta^2 \right)$. Then,

$$P\left(\|\mathbb{P}_n - P_Z\|_{\mathscr{F}} > \delta \right) \leq 2P\left(\left\| \mathbb{P}_n - \mathbb{P}_n' \right\|_{\mathscr{F}} > \frac{\delta}{2} \right) \tag{2.49}$$

Proof. For convenience in the notation, we assume that $\mathscr{F}$ is centered, i.e. $P_Z f = 0$ for all $f \in \mathscr{F}$. It holds that

$$\left\{ \|\mathbb{P}_n - P_Z\|_{\mathscr{F}} > \delta \right\} = \left\{ \exists f \in \mathscr{F} : \left| (\mathbb{P}_n - P_Z) f \right| > \delta \right\} =: A_{f^\star}.$$

Notice that f is a random function that depends only $\mathbb{Z}$. Therefore, we denote this function by $f_{\mathbb{Z}}^\star$. Notice that this function is fixed conditionally on $\mathbb{Z}$. From the law of iterated expectations, it follows

[12] We have used exactly the same argument in the final part of the proof of Theorem 2.3.3.

that

$$P\left(\left\{\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right\},\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}\right)=\mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\left\{\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right\}}\cdot\mathbb{1}_{\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}}\,\middle|\,\mathbb{Z}\right)\right]$$
$$=\mathbb{E}\left[\mathbb{1}_{\left\{\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right\}}\cdot\mathbb{E}\left(\mathbb{1}_{\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}}\,\middle|\,\mathbb{Z}\right)\right], \tag{2.50}$$

where the last step follows from the fact that $\mathbb{P}_n f_{\mathbb{Z}}^\star$ is $\mathbb{Z}$-measurable. From Equation (2.48) and from the conditions $n\geq 8\mathrm{Var}\left(f_{\mathbb{Z}}^\star(Z)\right)/\delta^2$ and $P_Z f_{\mathbb{Z}}^\star=0$ (conditionally on $\mathbb{Z}$), it follows that

$$\mathbb{E}\left(\mathbb{1}_{\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}}\,\middle|\,\mathbb{Z}\right)=P\left(\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\mid\mathbb{Z}\right)\geq 1/2.$$

It follows from Equation (2.50) that

$$P\left(\left\{\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right\},\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}\right)\geq\frac{1}{2}P\left(\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right).$$

Therefore,

$$\begin{aligned}P\left(\left\|\mathbb{P}_n-P_Z\right\|_{\mathscr{F}}>\delta\right)&\overset{P_Z f=0}{=}P\left(\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right)\\&\leq 2P\left(\left\{\left|\mathbb{P}_n f_{\mathbb{Z}}^\star\right|>\delta\right\},\left\{\left|\mathbb{P}'_n f_{\mathbb{Z}}^\star\right|\leq\delta/2\right\}\right)\\&\leq 2P\left(\left|\left(\mathbb{P}_n-\mathbb{P}'_n\right)f_{\mathbb{Z}}^\star\right|>\frac{\delta}{2}\right)\\&\leq 2P\left(\left\|\mathbb{P}_n-\mathbb{P}'_n\right\|_{\mathscr{F}}>\frac{\delta}{2}\right),\end{aligned} \tag{2.51}$$

where we used the triangle inequality in the third step. This last inequality finishes the proof. □

From the previous lemma, and from the observation that

$$\left\{f(Z_i)-f(Z'_i)\right\}_{i=1}^n\quad\text{and}\quad\left\{\varepsilon_i\left(f(Z_i)-f(Z'_i)\right)\right\}_{i=1}^n$$

have the same distribution, we can deduce that, if $n\geq\sup_{f\in\mathscr{F}}\left(8\mathrm{Var}\left(f(Z)\right)/\delta^2\right)$, then

$$\begin{aligned}P\left(\left\|\mathbb{P}_n-P_Z\right\|_{\mathscr{F}}>\delta\right)&\leq 2P\left(\left\|\mathbb{P}_n-\mathbb{P}'_n\right\|_{\mathscr{F}}>\frac{\delta}{2}\right)\\&=2P\left(\left\|\mathbb{P}_n^\varepsilon-\mathbb{P}_n'^{,\varepsilon}\right\|_{\mathscr{F}}>\frac{\delta}{2}\right)\\&\leq 2\left(P\left(\left\|\mathbb{P}_n^\varepsilon\right\|_{\mathscr{F}}>\frac{\delta}{4}\right)+P\left(\left\|\mathbb{P}_n'^{,\varepsilon}\right\|_{\mathscr{F}}>\frac{\delta}{4}\right)\right)\end{aligned}$$

$$= 4P\left(\|\mathbb{P}_n^\varepsilon\|_{\mathscr{F}} > \frac{\delta}{4}\right). \tag{2.52}$$

where we used the triangle inequality in the third step.

Theorem 2.5.7. Let $Z_1, \ldots, Z_n : \Omega \to \mathcal{Z}$ be i.i.d. random elements with distribution P_Z and let $\mathscr{F}$ be a collection of functions $\mathcal{Z} \to \mathbb{R}$. Suppose that:

- the class $\mathscr{F}$ is uniformly bounded by b.
- $\frac{1}{n} H_1(\delta, \mathscr{F}, \mathbb{P}_n) \xrightarrow{P} 0$ for all $\delta > 0$.

Then, $\mathscr{F}$ is a GC class.

Proof. Fix $\delta > 0$ and let $\{f_j\}_{j=1}^N$ be a (possibly infinite) $\delta/8$-covering set of $(\mathscr{F}, P_n)$, where $N = N_1(\delta, \mathscr{F}, P_n) < \infty$. This means that, for all $f \in \mathscr{F}$, there exists an index $j \in \{1, \ldots, N\}$, such that

$$\frac{1}{n}\sum_{i=1}^n |f(Z_i) - f_j(Z_i)| < \delta/8.$$

Recall that $\mathbb{P}_n^\varepsilon f := \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i)$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables, independent from $Z_1, \ldots, Z_n$. We can easily show that, for the index j described above, it also holds that $\left|\mathbb{P}_n^\varepsilon (f - f_j)\right| < \delta/8$. Therefore, by choosing the index j in a suitable way such that $\left|\mathbb{P}_n^\varepsilon (f - f_j)\right|$ is minimized, we can show that, for any $f \in \mathscr{F}$,

$$|\mathbb{P}_n^\varepsilon f| \le \min_{1\le j\le N} |\mathbb{P}_n^\varepsilon (f - f_j)| + \max_{1\le j\le N} |\mathbb{P}_n^\varepsilon f_j|,$$

which implies that

$$\sup_{f\in\mathscr{F}} |\mathbb{P}_n^\varepsilon f| \le \frac{\delta}{8} + \max_{1\le j\le N} |\mathbb{P}_n^\varepsilon f_j|.$$

Therefore,

$$\begin{aligned} P\left(\sup_{f\in\mathscr{F}} |\mathbb{P}_n^\varepsilon f| > \frac{\delta}{4}\right) &\le P\left(\max_{1\le j\le N} |\mathbb{P}_n^\varepsilon f_j| > \frac{\delta}{8}\right) \\ &\le P\left(\max_{1\le j\le N} |\mathbb{P}_n^\varepsilon f_j| > \frac{\delta}{8}\right). \end{aligned}$$

Using Hoeffding's inequality in combination with the union bound yields that, for all $t \ge 0$,

$$P\left(\max_{1\le j\le N} |\mathbb{P}_n^\varepsilon f_j| > b\sqrt{\frac{2\,(t + \log(2N))}{n}}\right) \le \exp(-t). \tag{2.53}$$

Notice that all these expressions still hold if $N = \infty$, because in this case the above bounds are trivially true. Consider the set

$$A_n := \left\{ \frac{\log\left(2N_1\left(\delta, \mathscr{F}, \mathbb{P}_n\right)\right)}{n} > \frac{\delta^2}{2^8 b^2} \right\}.$$

Also, choose $t = \frac{n\delta^2}{2^8 b^2}$. From the second condition of the theorem, it follows that $P\left(A_n\right) \xrightarrow{P} 0$. On the set A_n, it holds that

$$b\sqrt{\frac{2\left(t + \log(2N)\right)}{n}} > \frac{\delta}{8}$$

Combining the above argument, we deduce from Equation (2.52) that

$$\begin{aligned}
P\left(\left\|\mathbb{P}_n - P_Z\right\|_{\mathscr{F}} > \delta\right) &\leq 4P\left(\left\|\mathbb{P}_n^\varepsilon\right\|_{\mathscr{F}} > \frac{\delta}{4}\right) \\
&\leq P\left(\max_{1\leq j\leq N}\left|\mathbb{P}_n^\varepsilon f_j\right| > \frac{\delta}{8}\right) \\
&= 4\left(P\left(A_n \cap \left\{\max_{1\leq j\leq N}\left|\mathbb{P}_n^\varepsilon f_j\right| > \frac{\delta}{8}\right\}\right) + P\left(A_n^c \cap \left\{\max_{1\leq j\leq N}\left|\mathbb{P}_n^\varepsilon f_j\right| > \frac{\delta}{8}\right\}\right)\right) \\
&\leq 4\left(P\left(\max_{1\leq j\leq N}\left|\mathbb{P}_n^\varepsilon f_j\right| > b\sqrt{\frac{2\left(t + \log(2N)\right)}{n}}\right) + P\left(A_n^c\right)\right) \\
&\leq 4\left(\exp\left\{-\frac{n\delta^2}{2^8 b^2}\right\} + P\left(A_n^c\right)\right) \xrightarrow{P} 0,
\end{aligned}$$

which finishes the proof. □

Remark 2.5.8. The assumption that $\mathscr{F}$ is uniformly bounded can actually be replaced by a weaker one. For a function class $\mathscr{F}$, we define its *envelope* as

$$F = \sup_{f\in\mathscr{F}} |f|. \tag{2.54}$$

If $\mathscr{F}$ is uniformly bounded by b, then it holds that $F(z) \leq b$ for all $z \in \mathcal{Z}$. However, we can replace this assumption with the weaker condition that $P_Z F := \int_{\mathcal{Z}} F\, dP_Z < \infty$, and Theorem 2.5.7 continues to hold. The proof is relatively straightforward and can be found in [van de Geer, 2000, Theorem 3.7].

Before the end of this section, we present one example that illustrates how Theorem 2.5.7 can be used in practice. This example resembles Example 2.4.7, but it shows how much easier it is to derive the ULLN through $H_1\left(\delta, \mathscr{F}, \mathbb{P}_n\right)$ rather than $H_{1,B}\left(\delta, \mathscr{F}, P_Z\right)$

Example 2.5.9. Consider the class

$$\mathscr{F}_{\text{iso}} := \left\{f : \mathbb{R} \to [0,1] \mid f \nearrow \mathbb{R}\right\}.$$

We shall show that, for all $\delta > 0$, it holds that

$$H_\infty\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right) \leq \left(1 + \frac{1}{\delta}\right) \log\left(n + \frac{1}{\delta}\right), \tag{2.55}$$

where $\|f\|_{,\infty} := \max_{1 \leq i \leq n} |f(Z_i)|$. For any $p \in [1, \infty)$, it is easy to see that $\|f\|_{n,\infty} \geq \|f\|_{n,p}$, so

$$H_\infty\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right) > H_1\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right).$$

Thus, Equation (2.55) implies that $\mathscr{F}_{\text{iso}}$ is a GC class. To prove Equation (2.55), we first relabel the points $Z_1, \ldots, Z_n$ so that $Z_1 \leq \ldots \leq Z_n$. We also set $Z_0 = -\infty$ and $Z_{n+1} = +\infty$. We define the piecewise constant function

$$\widetilde{f} = \left\lfloor \frac{f(Z_1)}{\delta} \right\rfloor \cdot \delta \cdot \mathbb{1}_{(Z_0, Z_1)} + \sum_{i=1}^{n} \left\lfloor \frac{f(Z_i)}{\delta} \right\rfloor \cdot \delta \cdot \mathbb{1}_{[Z_i, Z_{i+1})}.$$

For all $i \in \{1, \ldots, n\}$, it holds that

$$\begin{aligned} f(Z_i) - \widetilde{f}(Z_i) &= f(Z_i) - \left\lfloor \frac{f(Z_i)}{\delta} \right\rfloor \cdot \delta \\ &\geq f(Z_i) - \frac{f(Z_i)}{\delta} \cdot \delta \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} f(Z_i) - \widetilde{f}(Z_i) &= f(Z_i) - \left\lfloor \frac{f(Z_i)}{\delta} \right\rfloor \cdot \delta \\ &\leq f(Z_i) - \left(\frac{f(Z_i)}{\delta} - 1 \right) \cdot \delta \\ &= \delta, \end{aligned}$$

so $\left\| f - \widetilde{f} \right\|_{n,\infty} \leq \delta$. It remains to compute the number of distinct piecewise constant functions that can be formed in this way.

This is a purely counting argument. Since f takes values in $[0, 1]$, these functions take values in the set $\Delta := \{0, \delta, \ldots, N\delta\}$, where $N \leq 1/\delta$. Their jumps belong to the set $\{Z_1, \ldots, Z_n\}$. However, we might have *multiple* jumps at the same point, e.g. if the function jumps directly from δ to 4δ. Even if we count the jumps together with their multiplicity (i.e. we count a triple jump as three different jumps), the total number of jumps cannot exceed N, since $\widetilde{f}$ takes values in Δ. Therefore, the number of distinct functions is upper bounded by the number of ways to choose n elements from a set of

cardinality $N+1$ (with replacement). The latter quantity is equal to

$$\binom{N+n}{n} \leq \binom{\lfloor 1/\delta \rfloor + n}{n} = \binom{\lfloor 1/\delta \rfloor + n}{\lfloor 1/\delta \rfloor}.$$

Therefore, from standard bounds on binomial coefficients, it follows that

$$\begin{aligned} H_\infty(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n) &\leq \log \binom{\lfloor 1/\delta \rfloor + n}{\lfloor 1/\delta \rfloor} \\ &\leq \log \left[(\lfloor 1/\delta \rfloor + n)^{\lfloor 1/\delta \rfloor} \right] \\ &= \left\lfloor \frac{1}{\delta} \right\rfloor \log \left(n + \left\lfloor \frac{1}{\delta} \right\rfloor \right), \end{aligned}$$

which finishes the proof.

This example shows how much more convenient it is to use Theorem 2.5.7 to prove that certain function classes are GC. The fact that we are interested in the empirical entropy (i.e. with respect to $\mathbb{P}_n$ instead of P_Z) allows us to construct covering sets to approximate functions only on $Z_1, \ldots, Z_n$ and not on their entire domain.

2.6 Vapnik-Chervonenkis Dimension

Vapnik-Chervonenkis (VC) theory, which was introduced by [Vapnik and Chervonenkis, 1971], led to a renewed interest in uniform laws of large numbers. In this section, we focus on the VC dimension. One of the most intriguing features of the VC dimension is that it depends solely on the function class and not on the underlying probability distribution P_Z of the random elements $Z_1, \ldots, Z_n$. This feature puts VC dimension in contrast to entropy and Rademacher dimension, which of course depend on P_Z or $\mathbb{P}_n$. For that reason, the VC dimension allows us to disentangle the complexity of the function class from the data distribution in the GC problem. These properties have sparked a very active interest in the VC dimension, especially in the area of Statistical Machine Learning.

Definition 2.6.1. Let $(\mathcal{Z}, \mathcal{G})$ be a measurable space and let $\mathscr{D}$ be a collection of subsets of $\mathcal{Z}$. For $z_1, \ldots, z_n \in \mathcal{Z}$, we define

$$\Delta^{\mathscr{D}}(z_1, \ldots, z_n) := \left| \{ D \cap \{z_1, \ldots, z_n\} : D \in \mathscr{D} \} \right|. \tag{2.56}$$

$\Delta^{\mathscr{D}}(z_1, \ldots, z_n)$ is commonly referred to as the number of subsets of $\{z_1, \ldots, z_n\}$ *shattered* by $\mathscr{D}$. This number is expected to grow as the size and complexity of $\mathscr{D}$ increase. For example, if $\mathscr{D} = 2^{\mathcal{Z}}$, then $\Delta^{\mathscr{D}}(z_1, \ldots, z_n) = 2^n$ *for any* $z_1, \ldots, z_n \in \mathcal{Z}$.

If, for some collection $\mathscr{D} \subset 2^{\mathcal{Z}}$ and some set of points $z_1, \ldots, z_n \in \mathcal{Z}$ it holds that $\Delta^{\mathscr{D}}(z_1, \ldots, z_n) = 2^n$, then we say that $\{z_1, \ldots, z_n\}$ is *shattered* by $\mathscr{D}$. In this case, $\mathscr{D}$ can pick out any of the 2^n subsets of $\{z_1, \ldots, z_n\}$. For any $\mathscr{D} \subset 2^{\mathcal{Z}}$, we define the maximal shattering number as

$$m^{\mathscr{D}}(n) := \sup\left\{\Delta^{\mathscr{D}}(z_1, \ldots, z_n) : z_1, \ldots, z_n \in \mathcal{Z}\right\}. \tag{2.57}$$

If $m^{\mathscr{D}}(n) = 2^n$ for some $n \geq 1$, then *there exists* a set of n points that is shattered by $\mathscr{D}$.

We are going to measure the complexity of $\mathscr{D}$ by looking at the largest possible size of a set that is shattered by it. This is the motivation for the definition of the VC dimension.

Definition 2.6.2 (VC dimension)**.** Using the above notation, we define the VC dimension of a collection $\mathscr{D} \subset 2^{\mathcal{Z}}$ by

$$V(\mathscr{D}) := \inf\left\{n \geq 1 : m^{\mathscr{D}}(n) < 2^n\right\}. \tag{2.58}$$

When $V(\mathscr{D}) < \infty$, we say that $\mathscr{D}$ is a VC class of sets.

In other words, the VC dimension of $\mathscr{D}$ is the integer $n \geq 1$ where $\mathscr{D}$ *fails* for the first time, in the sense that it is not able to shatter any set of size n. The following examples help us obtain some intuition on the VC dimension of a variety of classes of sets.

Example 2.6.3 (Half-intervals)**.** We consider the class $\mathscr{D} := \left\{\mathbb{1}_{(-\infty,z]} : z \in \mathbb{R}\right\}$. We shall show that $\Delta^{\mathscr{D}_1}(z_1, \ldots, z_n) \leq n+1$ for any set of distinct points $\{z_1, \ldots, z_n\}$. The idea is to use the order of $z_1, \ldots, z_n$ on the real line. Without loss of generality, we can assume that $z_1 < \ldots < z_n$. Then, $\mathscr{D}_1$ can only pick out the subsets $\emptyset, \{z_1\}, \{z_1, z_2\}, \ldots, \{z_1, \ldots, z_n\}$. Indeed, for any arbitrary $z \in \mathbb{R}$, there exist two cases:

- there exists an index $j \in \{1, \ldots, n-1\}$ such that $z_j < z \leq z_{j+1}$. Thus, $\{z_1, \ldots, z_n\} \cap (-\infty, z] = \{z_1, \ldots, z_j\}$.
- $z > z_n$, in which case $\{z_1, \ldots, z_n\} \cap (-\infty, z] = \{z_1, \ldots, z_n\}$.
- $z \leq z_1$, in which case we either have $\{z_1, \ldots, z_n\} \cap (-\infty, z] = \{z_1\}$ or $\{z_1, \ldots, z_n\} \cap (-\infty, z] = \emptyset$.

It follows that $\Delta^{\mathscr{D}_1}(z_1, \ldots, z_n) \leq n+1$. Strict inequality holds whenever at least two points z_i, z_j with $i \neq j$ coincide. It follows that $m^{\mathscr{D}_1}(n) \leq n+1$. Clearly, $\mathscr{D}_1$ shatters any singleton, but it cannot shatter any set with at least two distinct elements, since $n+1 < 2^n$ for $n \geq 2$. Therefore, $V(\mathscr{D}_1) = 2$. Using a similar argument, we can prove that the class $\mathscr{D}_d = \{(-\infty, t_1] \times \cdots \times (-\infty, t_d] : t_1, \ldots, t_d \in \mathbb{R}\}$

satisfies $m^{\mathscr{D}_d}(n) \leq (n+1)^d$. This grows polynomially in n, so $V(\mathscr{D}) < \infty$. This means that $\mathscr{D}_d$ is a VC class of sets.

Example 2.6.4 (Half-spaces)**.** We consider the class of halfspaces

$$\mathscr{D} := \left\{ \left\{ x \in \mathbb{R}^d \ : \theta^T x > w \right\} : \theta \in \mathbb{R}^d, w \in \mathbb{R} \right\}.$$

We shall show that $m^{\mathscr{D}}(n) \leq 2^d \binom{n}{d}$. Let $z_1, \ldots, z_n$ be any collection of points in $\mathbb{R}^d$. For any subset of d points, there exists a unique hyperplane containing them. This hyperplane partitions $\mathbb{R}^d$ into two halfspaces A and B. For the points that lie on the hyperplane, we can arbitrarily decide whether they belong to A or B in 2^d ways [13] For the rest of the points, it is clear whether they belong to A or B. Therefore, $\mathscr{D}$ can pick out at most $2^d \binom{n}{d}$ subsets. This still grows polynomially in n, which yields that $V(\mathscr{D}) < \infty$. In fact, [Pollard, 1984] showed that $V(\mathscr{D}) \leq d+2$, which is a much tighter bound than the one we obtain with the above argument.

The following lemma is useful when one wants to combine or transform VC classes of sets. Its proof is very straightforward and is omitted.

Lemma 2.6.5. Let $\mathscr{D}, \mathscr{E}$ be two VC classes of subsets of a space $\mathcal{Z}$. Then:

i. $\mathscr{D} \cup \mathscr{E}$ is a VC class.

ii. $\mathscr{D} \cap \mathscr{E}$ is a VC class.

iii. $\mathscr{D}^c := \{D^c : D \in \mathscr{D}\}$ is a VC class.

From the above definitions, one could imagine that, although $m^{\mathscr{D}}(n)$ might not be equal to 2^n for some value of n, it could still, in principle, take the value $2^n - 1$ - i.e. $\mathscr{D}$ could possibly fail to pick out just one subset. The following deep result, whose proof is omitted [14] shows that this cannot be the case. If $V(\mathscr{D}) < \infty$, then $m^{\mathscr{D}}(n)$ grows with a polynomial rate in n.

Theorem 2.6.6 (Sauer-Shelah Lemma)**.** If $V(\mathscr{D}) < \infty$, then it holds that

$$m^{\mathscr{D}}(n) \leq \sum_{i=0}^{V(\mathscr{D})} \binom{n}{i} \leq (n+1)^{V(\mathscr{D})}. \tag{2.59}$$

Although the quantities $\Delta^{\mathscr{D}}(z_1, \ldots, z_n), m^{\mathscr{D}}(n), V(\mathscr{D})$ seem to have a combinatorial nature, they can be used to derive uniform laws of large numbers.

[13] We can think of rotating the hyperplane slightly so that the points fall on one of the two halfspaces without affecting the position of the other points.

[14] The proof uses induction and it is very elementary but contains quite a few of technical details. It can be found in [Wainwright, 2019, Proposition 4.18].

Theorem 2.6.7 ([Vapnik and Chervonenkis, 1971]). Let $Z_1, \ldots, Z_n : \Omega \to \mathcal{Z}$ be i.i.d. random elements with distribution P_Z. If

$$\frac{1}{n} \log \Delta^{\mathscr{D}} (Z_1, \ldots, Z_n) \xrightarrow{P} 0, \tag{2.60}$$

then $\{\mathbb{1}_D : D \in \mathscr{D}\}$ is a GC function class.

Proof. Fix $\delta \in (0,1)$. Let $A_1, \ldots, A_\Delta$ be the (random) subsets of $\{Z_1, \ldots, Z_n\}$ that are picked out by $\mathscr{D}$, where $\Delta = \Delta^{\mathscr{D}} (Z_1, \ldots, Z_n)$. For each index $i \in \{1, \ldots, \Delta\}$, there exists a set $D_i \in \mathscr{D}$ such that $D_i \cap \{Z_1, \ldots, Z_n\} = A_i$. We consider the collection $\{\mathbb{1}_{D_i}\}_{i=1}^{\Delta}$. We show that this collection forms a δ-covering of $\{\mathbb{1}_D : D \in \mathscr{D}\}$. Indeed, fix $D \in \mathscr{D}$. Since $A_1, \ldots, A_\Delta$ is the complete lists of subsets of $\{Z_1, \ldots, Z_n\}$ picked out by $\mathscr{D}$, it follows that there exists an index $j \in \{1, \ldots, \Delta\}$ such that $D \cap \{Z_1, \ldots, Z_n\} = A_j$. This yields that

$$\left\| \mathbb{1}_D - \mathbb{1}_{D_j} \right\|_{1,n} = \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{1}_{D_j} (Z_i) - \mathbb{1}_D (Z_i) \right| = 0 < \delta,$$

because the intersections of D and D_j with $\{Z_1, \ldots, Z_n\}$ are exactly the same. This shows that $N_1 (\delta, \{\mathbb{1}_D : D \in \mathscr{D}\}, \mathbb{P}_n) \leq \Delta$. Thus, the given condition implies that

$$\frac{1}{n} H_1 (\delta, \{\mathbb{1}_D : D \in \mathscr{D}\}, \mathbb{P}_n) \xrightarrow{P} 0.$$

The result now follows from Theorem 2.5.7. □

From Theorem 2.6.6, we can deduce that any VC class of sets satisfies the condition

$$\frac{1}{n} \log \Delta^{\mathscr{D}} (Z_1, \ldots, Z_n) \xrightarrow{P} 0.$$

Thus, we obtain the following important result.

Lemma 2.6.8. If $\mathscr{D}$ is a VC class of sets, then $\{\mathbb{1}_D : D \in \mathscr{D}\}$ is a GC function class.

Perhaps it is surprising how straightforward it was to formulate a result about GC function classes using the newly introduced concept of VC dimension. However, this result only concerns classes of indicator functions. Providing similar results for general function classes requires some more work. VC dimension seems to be a concept that is inherently connected to sets, not functions. However, we can make the connection to function classes through the following definition.

Definition 2.6.9. Let $f : \mathcal{Z} \to \mathbb{R}$ be a function. The *graph* of f is defined as the set

$$\text{graph}\,(f) := \{(z,t) \in \mathcal{Z} \times \mathbb{R} : f(z) \geq t\}. \tag{2.61}$$

We say that a class $\mathscr{F}$ of functions is VC if $\{\text{graph}(f) : f \in \mathscr{F}\}$ is a VC class of sets.

For instance, Example 2.6.4 shows that the linear functions on $\mathbb{R}^d$ form a VC class. Out goal is to show that any VC class of functions must satisfy the ULLN. We start with the following technical lemma:

Lemma 2.6.10. Let N, v, C be constants such that $N \leq C\log^v(N)$. Then, it holds that

$$N \leq C\log^v\left(C^2\,(2v)^{2v}\right). \tag{2.62}$$

Proof. Using the inequality $\log x < x$, we obtain that $\log^v(N) = \left(2v\log\left(N^{\frac{1}{2v}}\right)\right)^v \leq (2v)^v N^{1/2}$. Thus, from the given condition, it follows that

$$\begin{aligned} N \leq C\,(2v)^v N^{1/2} &\Rightarrow N \leq C^2\,(2v)^{2v} \\ &\Rightarrow N \leq C\log^v\left(C^2\,(2v)^{2v}\right), \end{aligned}$$

where in the last step we used the given condition for the second time. □

Remark 2.6.11. The Sauer-Shelah lemma tells us that VC classes of sets satisfy $m^{\mathscr{D}}(n) \leq (n+1)^{V(\mathscr{D})}$. If $V(\mathscr{D})$ is fixed, then there exists a universal constant C such that $(n+1)^{V(\mathscr{D})} \leq C \cdot n^{V(\mathscr{D})}$ for all $n \geq 1$. Thus, we can characterize a VC class of sets as a class for which $m^{\mathscr{D}}(n) \leq Cn^{V(\mathscr{D})}$ for some class-specific constant C.

To show that a VC class of functions $\mathscr{F}$ satisfies the ULLN, we are going to bound the covering number of this function class using the VC dimension of $\mathscr{D}_{\mathscr{F}} := \{\text{graph}(f) : f \in \mathscr{F}\}$.

Lemma 2.6.12. Let Q be an arbitrary probability measure on $\mathcal{Z}$ and suppose that the class $\mathscr{F}$ is VC. Denote the VC dimension of $\mathscr{D}_{\mathscr{F}}$ by V. Fix $\delta > 0$ and let F be the envelope of $\mathscr{F}$, as it was defined in Equation 2.54. Then, there exists a constant $A = A(C, V)$ such that

$$N_1(\delta \cdot QF, \mathscr{F}, Q) \leq A\left(\frac{1}{\delta}\right)^V \log^V\left(\frac{1}{\delta}\right), \tag{2.63}$$

where C is the constant defined in Remark 2.6.11.

Proof. This is a rather indirect proof, and the result comes up in a surprising way. The argument uses a technique known as the *Probabilistic Method*, which is largely used in combinatorics.

Without loss of generality, we can assume that $QF = 1$. Otherwise, we simply replace δ by δ/QF and adjust the value of the constant A. Let $\{f_1, \ldots, f_M\}$ be a maximal δ-packing set. This means that, for all distinct indices $i, j \in \{1, \ldots, M\}$, it holds that $Q\left|f_i - f_j\right| > \delta$. We define a random element $S : \Omega \to \mathcal{Z}$ with distribution

$$P\left(S \in A\right) = \int_A F \, dQ, \; A \in \mathscr{G}.$$

Given $S = s$, we define a random variable $T \sim \text{Unif}\left[-F(s), F(s)\right]$. For all indices $k, j \in \{1, \ldots, M\}$, it holds that

$$P\left(T \text{ lies between } f_k(s), f_j(s) \;\middle|\; S = s\right) = \frac{\left|f_j(s) - f_k(s)\right|}{2F(s)}.$$

This implies that

$$P\left(T \text{ lies between } f_k(S), f_j(S)\right) = \int_{\mathcal{Z}} \frac{\left|f_j(s) - f_k(s)\right|}{2F(s)} \cdot F \, dQ > \frac{\delta}{2}.$$

Given i.i.d. pairs $\left(S_1, T_1\right), \ldots, \left(S_n, T_n\right)$, it follows that

$$P\Big(\forall i \in \{1, \ldots, n\}, \; T_i \text{ does not lie between } f_k(S_i), f_j(S_i)\Big) \leq \left(1 - \frac{\delta}{2}\right)^n.$$

The union bound yields that

$$\begin{aligned} P\left(\bigcup_{k \neq j} \Big\{\forall i \in \{1, \ldots, n\}, \; T_i \text{ does not lie between } f_k(S_i), f_j(S_i)\Big\}\right) &\leq \binom{M}{2}\left(1 - \frac{\delta}{2}\right)^n \\ &\leq \frac{M^2}{2} \cdot \exp\left\{-\frac{n\delta}{2}\right\}. \end{aligned}$$

It follows from that inequality that, for $n \geq \lceil 4 \log M/\delta \rceil$,

$$P\Big(\forall (j,k) \; \exists i : T_i \text{ lies between } f_k(S_i), f_j(S_i)\Big) \geq \frac{1}{2} > 0.$$

Therefore, there exists an element $\omega \in \Omega$ such that

$$\Big\{\forall (j,k) \; \exists i : T_i(\omega) \text{ lies between } f_k(S_i(\omega)), f_j(S_i(\omega))\Big\} \subset \Big\{\Delta^{\mathscr{D}_{\mathscr{F}}}\left((S_1(\omega), T_1(\omega)), \ldots, (S_n(\omega), T_n(\omega))\right) \geq M\Big\},$$

because the condition in the LHS yields that the graphs of f_k, f_j pick out a different subset of

$$\Big\{(S_1(\omega), T_1(\omega)), \ldots, (S_n(\omega), T_n(\omega))\Big\},$$

and this happens for all pairs (j,k). This yields that

$$M \leq \Delta^{\mathscr{D}_{\mathscr{F}}}\left((S_1(\omega), T_1(\omega)), \ldots, (S_n(\omega), T_n(\omega))\right) \leq Cn^V. \tag{2.64}$$

This holds for any n such that $n \geq \lceil 4 \log M/\delta \rceil$. For sufficiently small δ, it holds that $8 \log M/\delta < \lceil 4 \log M/\delta \rceil$, so we can assume that $n \leq 8 \log M/\delta$. Thus, from Equation (2.64), it follows that $M \leq C \cdot 8^V \left(\frac{1}{\delta}\right)^V \log^V(M)$. From Lemma 2.6.10 it follows that, there exists a constant $A = A(C,V)$ such that

$$M \leq A \left(\frac{1}{\delta}\right)^V \log^V\left(\frac{1}{\delta}\right).$$

The final result now follows from Lemma 2.5.3, since $N(\delta, \mathscr{F}, Q) \leq M(\delta, \mathscr{F}, Q) = M$. □

This lemma generalizes Theorem 2.6.7 to general function classes through the following result.

Theorem 2.6.13. Let $Z_1, \ldots, Z_n : \Omega \to \mathcal{Z}$ be i.i.d. random elements with distribution P_Z, and let $\mathscr{F}$ be a function class with envelope F. If $P_Z F < \infty$ and $\mathscr{F}$ is a VC class, then $\mathscr{F}$ is also a GC class.

Proof. From Lemma 2.6.12, it follows that $N_1(\delta, \mathscr{F}, \mathbb{P}_n) \leq N(\delta/\mathbb{P}_n F)$, where

$$N(\delta) := A \left(\frac{1}{\delta}\right)^V \log^V\left(\frac{1}{\delta}\right).$$

$N(\delta)$ is a strictly decreasing function of δ. Since $P_Z F < \infty$, it follows from the SLLN that

$$P(\mathbb{P}_n F > 2P_Z F) \xrightarrow{P} 0.$$

Therefore,

$$\begin{aligned} P\left(N_1(\delta, \mathscr{F}, \mathbb{P}_n) > N\left(\frac{\delta}{2P_Z F}\right)\right) &\leq P\left(N\left(\frac{\delta}{\mathbb{P}_n F}\right) > N\left(\frac{\delta}{2P_Z F}\right)\right) \\ &= P(\mathbb{P}_n F > 2P_Z F) \\ &\xrightarrow{P} 0. \end{aligned}$$

The final result now follows directly from Theorem 2.5.7. □

2.7 Consistency of ERM Estimators

In this section, we revisit Empirical Risk Minimization and use the results that we have presented to prove that several common Empirical Risk Minimizers are consistent. What does consistency mean in this setting? Recall from Example 2.1.1 that our goal is to minimize the expected risk $R\left(f, f^{\star}\right) := P_{f^{\star}}\mathcal{L}_f\left(Z\right) = \mathbb{E}\left[\mathcal{L}_f\left(Z\right)\right]$, where $f^{\star}$ is the true value of the parameter $f \in \mathscr{F}$. Therefore, we should be looking for

$$\arg\min_{f\in\mathscr{F}} R\left(f, f^{\star}\right).$$

However, since $f^{\star}$ is unknown, we resort to the empirical risk $\widehat{R}_n\left(f\right) := \mathbb{P}_n\mathcal{L}_f = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_f\left(Z_i\right)$ and we use the estimator

$$\widehat{f}_n \in \arg\min_{f\in\mathscr{F}} \widehat{R}_n\left(f\right).$$

However, since our true goal is to minimize $R\left(f, f^{\star}\right)$ and not $\widehat{R}\left(f\right)$, we would still hope that $\widehat{f}_n$ performs well in terms of $R\left(f, f^{\star}\right)$, namely that

$$R\left(\widehat{f}_n, f^{\star}\right) \approx \inf_{f\in\mathscr{F}} R\left(f, f^{\star}\right)$$

This is exactly the *consistency* property that we are interested in. The first lemma the we are going to show makes the connection of consistency with the ULLN.

Lemma 2.7.1. Suppose that the function class $\mathcal{L}_{\mathscr{F}} := \left\{\mathcal{L}_f : f \in \mathscr{F}\right\}$ is GC. If $f_0 \in \mathscr{F}$ minimizes $R\left(f, f^{\star}\right)$, then, it holds that

$$R\left(\widehat{f}_n, f^{\star}\right) \xrightarrow{P} R\left(f_0, f^{\star}\right).$$

Proof. This follows directly from Equation (2.14). Based on the discussion after this equation in Section 2.1, it follows that the condition

$$\sup_{f\in\mathscr{F}}\left|\widehat{R}_n(f) - R\left(f, f^{\star}\right)\right| \xrightarrow{P} 0 \text{ as } n \to \infty$$

is sufficient for the proof of the convergence $R\left(\widehat{f}_n, f^{\star}\right) \xrightarrow{P} R\left(f_0, f^{\star}\right)$. However, this condition is simply the ULLN for the class $\mathcal{L}_{\mathscr{F}}$, which is satisfied due to our assumption that this class is GC. □

Example 2.7.2 (Linear classification)**.** In Example 2.1.1, we looked at binary classification. One type of binary classification is the so called *linear* (binary) classification. In linear classification, we assume that the points $Z = (X, Y) \in \mathbb{R}^d \times \{0, 1\}$ are separated by a hyperplane, which determines the value of Y. Our goal is to estimate this hyperplane. There are several ways to do that, including

logistic regression and support vector machines. However, these approaches pare based on ERM and produce only an estimate of the true hyperplane. For example, in Figure 2.1, if we only have access to the large data points, we can choose the green line as the separating hyperplane. However,

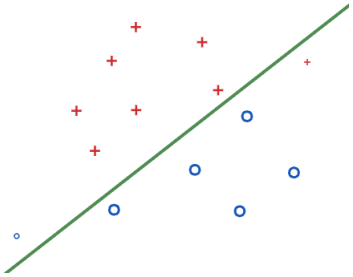

FIGURE 2.1

if we sample new data points from the underlying distribution (like the small ones on the sides), we might also encounter some incorrect classifications. Using the same notation as in Example 2.1.1, our assumption is that the true classification function $f^\star$ has the form

$$y = f^\star(x) = \mathbb{1}_{\{\theta_\star^\top x > w_\star\}}.$$

The underlying function class is $\mathscr{F} := \left\{\mathbb{1}_{\{\theta^\top x > w\}} : \theta \in \mathbb{R}^d, w \in \mathbb{R}\right\}$, and

$$\mathcal{L}_\mathscr{F} := \left\{\left|y - \mathbb{1}_{\{\theta^\top x > w\}}\right| : \theta \in \mathbb{R}^d, w \in \mathbb{R}\right\}.$$

For any function $f = \mathbb{1}_{\{\theta^\top x > w\}} \in \mathscr{F}$, it holds that $P_Z f = P\left(f(X) \neq Y\right)$, so we would like to find the values of $\theta \in \mathbb{R}^d, w \in \mathbb{R}$ that minimize the probability of misclassification. In this setting, where there is no noise, this hyperplane is clearly the true hyperplane determined by $\theta_\star, w_\star$ [15]. We can show that the hyperplane $\widehat{f}_n(x) = \mathbb{1}_{\{\widehat{\theta}_n^\top x > \widehat{w}_n\}}$ that we derive from ERM is consistent, in the sense that

$$P\left(\widehat{f}_n(X) \neq Y\right) \xrightarrow{P} 0 = P\left(f^\star(X) \neq Y\right).$$

We first show that $\mathcal{L}_\mathscr{F}$ is a VC function class. Notice that

$$\mathcal{L}_\mathscr{F} := \left\{\mathbb{1}_{\{\theta^\top x > w\} \triangle \{\theta_\star^\top x > w_\star\}} : \theta \in \mathbb{R}^d, w \in \mathbb{R}\right\}.$$

The class of sets $\mathscr{D} := \left\{\{\theta^\top x > w\} : \theta \in \mathbb{R}^d, w \in \mathbb{R}\right\}$ is VC due to Example 2.6.4. It follows from Lemma 2.6.5 that

$$\mathscr{D} \triangle \mathscr{D} := \{D_1 \triangle D_2 : D_1, D_2 \in \mathscr{D}\}$$

[15] The same actually holds in a noisy setting, under some assumptions on the noise variables.

is also a VC class of sets. The class

$$\left\{D \triangle \left\{\theta_\star^\top x > w_\star\right\} : D \in \mathscr{D}\right\}$$

is a subclass of $\mathscr{D} \triangle \mathscr{D}$, so it is also a VC class. It follows now from Lemma 2.6.8 that $\mathcal{L}_\mathscr{F}$ is a VC class of functions. Since $\mathcal{L}_\mathscr{F}$ is a VC function class, it follows from Theorem 2.6.13 that it is also a GC class. Finally, it follows from Lemma 2.7.1 that

$$R\left(\widehat{f}_n, f^\star\right) \xrightarrow{P} R\left(f_0, f^\star\right).$$

In our setting, it holds that $R\left(f_0, f^\star\right) = R\left(f^\star, f^\star\right) = 0$, which yields that

$$P\left(\widehat{f}_n(X) \neq Y\right) \xrightarrow{P} 0.$$

Example 2.7.3 (Isotonic regression)**.** Suppose that $Z = (X, Y) \in \mathbb{R} \times \mathbb{R}$ and that $Y = f^\star(X) + \xi$, where $f^\star : \mathbb{R} \to [0,1]$ is an increasing function and ξ is a random variable that is independent of X, such that $\mathbb{E}\xi = 0$, $\mathrm{Var}(\xi) = \sigma^2 < \infty$, and $\mathbb{E}\xi^4 < \infty$. The true function $f^\star$ is unknown and the only known information is that it belongs to the class

$$\mathscr{F}_{\text{iso}} := \{f : \mathbb{R} \to [0,1] : f \text{ is increasing}\}.$$

The loss function incurred by a point $f \in \mathscr{F}_{\text{iso}}$ is the squared error: $\mathcal{L}_f(x,y) := (y - f(x))^2$. Thus,

$$\mathcal{L}_{\mathscr{F}_{\text{iso}}} = \left\{\mathcal{L}_f(x,y) : f \in \mathscr{F}_{\text{iso}}\right\} = \left\{(y - f(x))^2 : f \in \mathscr{F}_{\text{iso}}\right\}.$$

The empirical minimizer is any element

$$\widehat{f}_n \in \arg\min_{f \in \mathscr{F}_{\text{iso}}} \left(\frac{1}{n}\sum_{i=1}^n \left(Y_i - f\left(X_i\right)\right)^2\right).$$

To show that this is a GC class, we use the following lemma:

Lemma 2.7.4. Let $\mathscr{F}$ be a function class on a space $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$. Suppose that $Z = (X, Y) \in \mathcal{Z}$ is a pair of random elements such that $Y = f^\star(X) + \xi$, where $f^\star \in \mathscr{F}$ and ξ is a zero-mean noise variable with variance $\sigma^2 < \infty$. If:

- $\frac{1}{n} \log N_2\left(\delta, \mathscr{F}, \mathbb{P}_n\right) \xrightarrow{P} 0$,
- the envelope of $\mathcal{L}_\mathscr{F}$ is an element of $L_2\left(P_Z\right)$,

then $\mathcal{L}_{\mathscr{F}}$ is a GC class, where $\mathcal{L}_f$ is the squared loss function.

Proof. This result relies on Theorem 2.5.7- We first use the Cauchy-Schwarz inequality to show that the first given condition reduces to the second condition of that theorem. For any $f, g \in \mathscr{F}$, it holds that

$$\begin{aligned}
\mathbb{P}_n \left|\mathcal{L}_f - \mathcal{L}_g\right| &= \frac{1}{n}\sum_{i=1}^{n}\left|\left(Y_i - f\left(X_i\right)\right)^2 - \left(Y_i - g\left(X_i\right)\right)^2\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|\left(\xi_i + \left(f^\star - f\right)\left(X_i\right)\right)^2 - \left(\xi_i + \left(f^\star - g\right)\left(X_i\right)\right)^2\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|2\xi_i\left(g\left(X_i\right) - f\left(X_i\right)\right) + \left(f^\star - f\right)^2\left(X_i\right) - \left(f^\star - g\right)^2\left(X_i\right)\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|2\xi_i\left(g\left(X_i\right) - f\left(X_i\right)\right) + \left(g\left(X_i\right) - f\left(X_i\right)\right)\left(2f^\star\left(X_i\right) - f\left(X_i\right) - g\left(X_i\right)\right)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[4\xi_i^2 + \left(2\left|f^\star\right| + |f| + |g|\right)^2\left(X_i\right)\right]} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(g(X_i) - f(X_i)\right)^2} \\
&= \underbrace{\sqrt{\frac{4}{n}\sum_{i=1}^{n}\xi_i^2 + \mathbb{P}_n\Big(2\left|f^\star\right| + |f| + |g|\Big)^2}}_{\Delta_n} \cdot \sqrt{\mathbb{P}_n\left(g - f\right)^2}.
\end{aligned} \tag{2.65}$$

It follows that, for all $\delta > 0$,

$$N_1\left(\delta, \mathcal{L}_{\mathscr{F}}, \mathbb{P}_n\right) \leq N_2\left(\delta \cdot \Delta_n, \mathscr{F}, \mathbb{P}_n\right). \tag{2.66}$$

The SLLN yields that

$$\mathbb{P}_n\Big(2\left|f^\star\right| + |f| + |g|\Big)^2 \xrightarrow{P} P_Z\Big(2\left|f^\star\right| + |f| + |g|\Big)^2 \leq 16 P_Z F < \infty$$

and

$$\frac{4}{n}\sum_{i=1}^{n}\xi_i^2 \xrightarrow{P} 4\sigma^2.$$

This shows that $\Delta_n \xrightarrow{P} 64\sigma^2 P_Z F =: \Delta < \infty$. For all $\varepsilon > 0$, we denote $\left\{\frac{1}{n}\log N_1\left(\delta, \mathcal{L}_{\mathscr{F}}, \mathbb{P}_n\right) > \varepsilon\right\}$ by $A_{n,\varepsilon}$. It follows from Equation (2.66) that,

$$\begin{aligned}
P\left(A_{n,\varepsilon}\right) &= P\left(A_{n,\varepsilon} \cap \left\{\Delta_n < \Delta/2\right\}\right) + P\left(A_{n,\varepsilon} \cap \left\{\Delta_n \geq \Delta/2\right\}\right) \\
&\leq P\left(\Delta_n < \Delta/2\right) + P\Big(\frac{1}{n}\log N_2\left(\Delta\delta/2, \mathscr{F}, \mathbb{P}_n\right) > \varepsilon\Big) \\
&\xrightarrow{P} 0,
\end{aligned}$$

where, in the last step, we used the fact that $\Delta_n \xrightarrow{P} \Delta$ and the first condition of the Lemma. This shows that the conditions of Theorem 2.5.7 hold, so $\mathcal{L}_{\mathscr{F}}$ is a GC class. □

We now try to apply the result of this lemma for the class $\mathscr{F}_{\text{iso}}$.

- In Example 2.5.9, we showed that

$$H_\infty\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right) \leq \left\lfloor \frac{1}{\delta} \right\rfloor \log\left(n + \left\lfloor \frac{1}{\delta} \right\rfloor\right).$$

  For all $\delta > 0$ it holds that $N_2\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right) \leq N_\infty\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right)$, so, from the above bound, it follows that $\frac{1}{n} \log N_2\left(\delta, \mathscr{F}_{\text{iso}}, \mathbb{P}_n\right) \xrightarrow{P} 0$.

- The class $\mathscr{F}_{\text{iso}}$ is uniformly bounded between 0 and 1, so, for all $f \in \mathscr{F}_{\text{iso}}$, it holds that

$$\begin{aligned}
\mathcal{L}_f(x, y) &= (y - f(x))^2 \\
&\leq \max\left\{(y-1)^2, y^2\right\} \\
&\leq (y-1)^2 + y^2.
\end{aligned}$$

  Therefore, for the envelope $\mathcal{L}_F$ of $\mathcal{L}_{\mathscr{F}}$, it holds that $F(x, y) \leq (y-1)^2 + y^2$. We have

$$P_Z\left[(Y-1)^2 + Y^2\right]^2 = P_Z\left[(f^\star(X) + \xi - 1)^2 + (f^\star(X) + \xi)^2\right]^2.$$

  Because of the assumption $\mathbb{E}\xi^4 < \infty$, and because $f^\star$ is bounded between 0 and 1, it follows that the above quantity is finite. This yields that $\mathcal{L}_F \in L_2(P_Z)$.

Indeed, from Lemma 2.7.4, it follows that $\mathcal{L}_{\mathscr{F}_{\text{iso}}}$ is a GC class, so, from Lemma 2.7.1, it follows that $R\left(\widehat{f}_n, f^\star\right) \xrightarrow{P} \inf_{f \in \mathscr{F}_{\text{iso}}} R(f, f^\star)$, which is what we wanted to show.

However, we can say a little bit more about that infimum. For all $f \in \mathscr{F}_{\text{iso}}$, it holds that

$$\begin{aligned}
R(f, f^\star) &= \mathbb{E}\left[\mathcal{L}_f(X, Y)\right] \\
&= \mathbb{E}\left(Y - f^\star(X)\right)^2 \\
&= \mathbb{E}\left(\xi - (f - f^\star)(X)\right)^2 \\
&= \sigma^2 + \mathbb{E}\left(f^\star(X) - f(X)\right)^2 - 2\mathbb{E}\left[\xi\left(f(X) - f^\star(X)\right)\right]. \qquad (2.67)
\end{aligned}$$

It follows from the independence of ξ, X and from $\mathbb{E}\xi = 0$ that the third term is equal to zero. Therefore, the theoretical risk is minimized by $f^\star$, or any other function that is almost surely equal to it. From Example 2.5.9, it follows that $\mathscr{F}_{\text{iso}}$ is a GC class, so Lemma 2.7.1 and Equation (2.67) yield

that

$$R\left(\widehat{f}_n, f^\star\right) \xrightarrow{P} \inf_{f \in \mathscr{F}_{\text{iso}}} R\left(f, f^\star\right) = \sigma^2.$$

Moreover, Equation (2.67) shows that $\mathbb{E}_X\left(f^\star(X) - \widehat{f}_n\left(X\right)\right)^2 \xrightarrow{P} 0$.

Example 2.7.5 (Changepoint detection)**.** We assume again that $Z = (X, Y) \in [0,1] \times \mathbb{R}$ and that $Y = f^\star(X) + \xi$, where $f^\star$ is an unknown function that now belongs to the class

$$\mathscr{F} := \left\{ f(x) = (a + bx)\, \mathbb{1}_{\{0 \leq x \leq c\}} + (d + ex)\, \mathbb{1}_{\{1 \geq x > c\}} : a, b, c, d, e \in [0,1] \right\}.$$

In other words, $f^\star$ is a piecewise linear function with an unknown changepoint. Like before, we use the squared loss function and we make the same assumptions about the noise variable ξ. In Example 2.4.7, we showed that, for all $\delta > 0$, it holds that $H_\infty\left(\delta, \mathscr{F}_{\text{Lipschitz}}\right) \leq A/\delta$. Notice that $\mathscr{F} \subset \mathscr{F}_{\text{Lipschitz}}$, so $H_\infty\left(\delta, \mathscr{F}\right) \leq A/\delta$. It also holds that

$$|f|_{n,\infty} := \max_{1 \leq i \leq n} |f(X_i)| \leq \sup_{x \in [0,1]} |f(x)| = |f|_\infty,$$

so $H_\infty\left(\delta, \mathscr{F}, \mathbb{P}_n\right) \leq H_\infty\left(\delta, \mathscr{F}\right)$. Combining these results, we obtain that

$$\frac{1}{n} \log N_2\left(\delta, \mathscr{F}, \mathbb{P}_n\right) \leq \frac{1}{n} H_\infty\left(\delta, \mathscr{F}, \mathbb{P}_n\right) \xrightarrow{P} 0.$$

Also, with an argument identical to the one in Example 2.7.3, we can show that $\mathcal{L}_{\mathscr{F}}$ has an L_2-integrable envelope. Lemma 2.7.4 yields that $\mathcal{L}_{\mathscr{F}}$ is a GC class, which implies that

$$R\left(\widehat{f}_n, f^\star\right) \xrightarrow{P} \inf_{f \in \mathscr{F}} R\left(f, f^\star\right) = R\left(f^\star, f^\star\right) = \sigma^2.$$

In the above examples, we used Lemma 2.7.4, as well as some set-theoretic arguments, to prove that the class $\mathcal{L}_{\mathscr{F}}$ is GC. However, there also exist some more generic ways to show this property. More specifically, it is possible to show [van de Geer, 2000, Lemma 3.1] that, if the envelope of $\mathcal{L}_{\mathscr{F}}$ belongs to $L_q(P_Z)$, the space $(\mathscr{F}, d)$ is compact, and $f \mapsto \mathcal{L}_f(z)$ is continuous for all $z \in \mathcal{Z}$, then $H_{q,B}\left(\delta, \mathcal{L}_{\mathscr{F}}, \mathbb{P}_n\right) < \infty$, so $\mathcal{L}_{\mathscr{F}}$ is a GC class.

2.8 Rates of Convergence of M-Estimators

In the previous section, the property that we were interested in was

$$R\left(\widehat{f}_n, f^\star\right) \xrightarrow{P} \inf_{f\in\mathscr{F}} R\left(f, f^\star\right).$$

We called this property *consistency*. However, if $\mathscr{F}$ is itself a (semi-)metric space, we can even talk about consistency in the traditional statistical sense, namely

$$\widehat{f}_n \xrightarrow{P} f^\star.$$

In this section, we equip $\mathscr{F}$ with the random semimetric $\|\cdot\|_n$ and investigate consistency from this classical perspective. On the way, we introduce *chaining*, a technique developed by Kolmogorov and brought to the spotlight of Empirical Process Theory by Richard Dudley.

Throughout this section, we focus on nonparametric regression. We assume that $\{x_i\}_{i=1}^n$ is a n-tuple of fixed points and that, for all $i \in \{1,\dots,n\}$,

$$Y_i = f^\star(x_i) + \varepsilon_i, \tag{2.68}$$

where $f^\star \in \mathscr{F}$ is an unknown function, and $\varepsilon_1,\dots,\varepsilon_n \sim \mathcal{N}\left(0,\sigma^2\right)$ are independent error terms. We use the squared loss function $\mathcal{L}_f(x,y) = (y - f(x))^2$. We can use the same argument as in Equation (2.67) to show that

$$f^\star = \arg\min_{f\in\mathscr{F}} P_Z \mathcal{L}_f.$$

On the other hand, the empirical risk minimizer is the estimator

$$\widehat{f}_n \in \arg\min_{f\in\mathscr{F}} \sum_{i=1}^n \left(Y_i - g(x_i)\right)^2.$$

The results we are going to present also generalize to other loss functions, and non-Gaussian errors. They also generalize to random design $\{X_i\}_{i=1}^n$ by considering the conditional measures induced by $X_1,\dots,X_n$. However, the simpler framework we introduced above does not lack anything in terms of intuition and technical tools, and is an adequate prototype for the more general theory. Therefore, we will stick to this framework and make comments about possible generalizations wherever these generalizations present theoretical interest.

One of the most fundamental tools in this section is the following inequality.

Lemma 2.8.1. Let $Z_1, \ldots, Z_n : \Omega \to \mathcal{Z}$ be i.i.d. random elements with distribution P_Z and let $\mathbb{P}_n$ be the induced empirical measure. Then,

$$0 \leq P_Z \left(\mathcal{L}_{\widehat{f}_n} - \mathcal{L}_{f^\star} \right) \leq - \left(\mathbb{P}_n - P_Z \right) \left(\mathcal{L}_{\widehat{f}_n} - \mathcal{L}_{f^\star} \right) \tag{2.69}$$

Proof. The inequality on the left simply uses the fact that

$$f^\star = \arg\min_{f \in \mathscr{F}} P_Z \mathcal{L}_f.$$

The one on the right can be written as $\mathbb{P}_n \left(\mathcal{L}_{\widehat{f}_n} - \mathcal{L}_{f^\star} \right) \leq 0$, which holds due to the fact that

$$\widehat{f}_n \in \arg\min_{f \in \mathscr{F}} \mathbb{P}_n \mathcal{L}_f.$$

□

The most suitable (semi-)norm for the investigation of the consistency of $\widehat{f}_n$ turns out to be $\|\cdot\|_{n,p}$, defined in Equation (2.45). Since we will only be dealing with the case $p = 2$, we denote $\|\cdot\|_{n,2}$ simply by $\|\cdot\|_n$. Also, since this norm is only determined by the values of the functions at $x_1, \ldots, x_n$, we identify any function $f \in \mathscr{F}$ with the vector

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}.$$

Under the model (2.68), the inequality in Lemma 2.8.1 takes the following form.

Lemma 2.8.2. Let us denote the vector $(\varepsilon_1, \ldots, \varepsilon_n)^\top$ by ε. Then, under the above notation, it holds that

$$\left\| \widehat{f}_n - f^\star \right\|_n \leq \frac{2}{n} \varepsilon^\top \left(\widehat{f}_n - f^\star \right). \tag{2.70}$$

Proof. For all $f \in \mathscr{F}$, it holds that

$$\begin{aligned} \mathbb{P}_n \mathcal{L}_f &= \sum_{i=1}^n \left(Y_i - f(x_i) \right)^2 \\ &= \|Y - f\|_n^2 \\ &= \|\varepsilon + (f^\star - f)\|_n^2 \\ &= \|\varepsilon\|_n^2 - \frac{2}{n} \varepsilon^\top (f^\star - f) + \|f^\star - f\|_n^2. \end{aligned}$$

From this Equation, it follows that

$$\mathbb{P}_n\left(\mathcal{L}_{\widehat{f}_n} - \mathcal{L}_f^\star\right) = -\frac{2}{n}\varepsilon^\top\left(\widehat{f}_n - f^\star\right) + \left\|\widehat{f}_n - f^\star\right\|_n^2.$$

Since $\widehat{f}_n$ minimizes the empirical risk, we obtain that

$$0 \geq -\frac{2}{n}\varepsilon^\top\left(\widehat{f}_n - f^\star\right) + \left\|\widehat{f}_n - f^\star\right\|_n^2,$$

which finishes the proof. □

For any $\delta > 0$, we denote the ball $\left\{f \in \mathscr{F} : \|f - f^\star\|_n \leq \delta\right\}$ by $\mathscr{F}(\delta)$. We will show that consistency in terms of $\|\cdot\|_n$ is very closely related to the empirical entropy $H_2(\cdot, \mathscr{F}(\delta), \mathbb{P}_n)$. For any $\delta > 0$, we define the entropy integral

$$J(\delta) := 2\int_0^\delta \sqrt{2H_2(u, \mathscr{F}(\delta), \mathbb{P}_n)}\, du. \tag{2.71}$$

For our purposes, the constant 2 could be removed from both instances, but it is kept for historical reasons. The proof of the following lemma is one of the most technical in this thesis, but it opens the path for the derivation of convergence rates for least squares estimators.

Lemma 2.8.3. For all $\delta > 0$ and $t > 0$, it holds with probability at least $1 - 2e^{-t}$ that

$$\sup_{f \in \mathscr{F}(\delta)}\left[\frac{1}{\sqrt{n}}\varepsilon^\top(f - f^\star)\right] \leq 2J(\delta) + 4\delta\sqrt{1+t} \tag{2.72}$$

Proof. This proof uses the method of *chaining*.

Fix $\delta > 0$ and $t > 0$. Then, for any integer $S \geq 1$, let $\left\{f_j^S\right\}_{j=1}^{N_S} \subset \mathscr{F}(\delta)$ be a minimal $2^{-S}\delta$-covering of $\mathscr{F}(\delta)$ with respect to $\|\cdot\|_n$. This means that $N_S = N_2\left(2^{-S}\delta, \mathscr{F}(\delta), \mathbb{P}_n\right)$.

For any $f \in \mathscr{F}(\delta)$, there exists an element $f^{S+1} \in \left\{f_j^{S+1}\right\}_{j=1}^{N_{S+1}}$ such that $\left\|f - f^{S+1}\right\|_n \leq 2^{-(S+1)}\delta$. Similarly, for the point $f^{S+1} \in \mathscr{F}(\delta)$, we can choose an element $f^S \in \left\{f_j^S\right\}_{j=1}^{N_S}$ such that

$$\left\|f^{S+1} - f^S\right\|_n = \min_{1 \leq k \leq N_S}\left\|f^{S+1} - f_j^S\right\|_n \leq 2^{-S}\delta.$$

We follow the same process for all $s \in \{1, \ldots, S-1\}$ by defining $f^s \in \left\{f_j^s\right\}_{j=1}^{N_S}$ recursively, in such a way that

$$\left\|f^{s+1} - f^s\right\|_n = \min_{1 \leq k \leq N_s}\left\|f^{s+1} - f_j^s\right\|_n \leq 2^{-s}\delta.$$

This way, we have defined a *chain*

$$f^{S+1}, f^S, \ldots, f^1.$$

After the choice of f^{S+1}, there is a unique choice of $f^S, \ldots, f^1$. Thus, the number of these chains is $N_2\left(2^{-(S+1)}\delta, \mathscr{F}(\delta), \mathbb{P}_n\right)$.

Since $\left\|f^1 - f^\star\right\|_n \leq \delta$, we can define $f^0 = f^\star$. We can now decompose $f - f^\star$ into the telescoping sum

$$f - f^\star = \left(f - f^{S+1}\right) + \sum_{s=0}^{S}\left(f^{s+1} - f^s\right).$$

It follows that

$$\frac{1}{\sqrt{n}}\varepsilon^\top (f - f^\star) = \frac{1}{\sqrt{n}}\varepsilon^\top \left(f - f^{S+1}\right) + \sum_{s=0}^{S}\frac{1}{\sqrt{n}}\varepsilon^\top \left(f^{s+1} - f^s\right) \tag{2.73}$$

Notice that

$$\frac{1}{\sqrt{n}}\varepsilon^\top \left(f^{s+1} - f^s\right) = \sum_{i=1}^{n}\frac{1}{\sqrt{n}}\left(f^{s+1}(X_i) - f^s(X_i)\right)\varepsilon_i,$$

which is a sum of independent centered Gaussian random variables with variances

$$\frac{1}{n}\left(f^{s+1}(X_1) - f^s(X_1)\right)^2, \ldots, \frac{1}{n}\left(f^{s+1}(X_n) - f^s(X_n)\right)^2$$

respectively. From Hoeffding's inequality,

$$\begin{aligned} P\left(\left|\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\left(f^{s+1}(X_i) - f^s(X_i)\right)\varepsilon_i\right| \geq 2^{-s}\delta\sqrt{2t}\right) &\leq 2\exp\left\{-\frac{2^{-2s}\delta^2\cdot 2t}{2\sum_{i=1}^{n}\frac{1}{n}\left(f^{s+1}(X_i) - f^s(X_i)\right)^2}\right\} \\ &= 2\exp\left\{-\frac{2^{-2s}\delta^2\cdot 2t}{\left\|f^{s+1} - f^s\right\|_n^2}\right\} \\ &\leq 2\exp\left\{-t\right\}, \end{aligned}$$

where the last inequality follows from the fact that $\left\|f^{s+1} - f^s\right\|_n \leq 2^{-s}\delta$. The union bound over all possible $N_2\left(2^{-(s+1)}\delta, \mathscr{F}(\delta), \mathbb{P}_n\right)$ chains $f^{s+1}, \ldots, f^0$ that can be induced by all elements of $\mathscr{F}(\delta)$ yields that

$$P\left(\sup_{f\in\mathscr{F}(\delta)}\left|\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\left(f^{s+1}(X_i) - f^s(X_i)\right)\varepsilon_i\right| \geq 2^{-s}\delta\sqrt{2\left(t + H_{s+1}\right)}\right) \leq 2\exp\left\{-t\right\}, \tag{2.74}$$

where $H_{S+1} := \log N_2\left(2^{-(S+1)}\delta, \mathscr{F}(\delta), \mathbb{P}_n\right)$. We deduce that

$$\begin{aligned}
&P\left(\sup_{f\in\mathscr{F}(\delta)}\left[\sum_{s=0}^{S}\frac{1}{\sqrt{n}}\varepsilon^\top\left(f^{s+1}-f^s\right)\right] \geq \sum_{s=0}^{S}2^{-s}\delta\sqrt{2\left((s+1)(t+1)+H_{s+1}\right)}\right)\\
&\leq \sum_{s=0}^{S}P\left(\sup_{f\in\mathscr{F}}\left[\frac{1}{\sqrt{n}}\varepsilon^\top\left(f^{s+1}-f^s\right)\right] \geq 2^{-s}\delta\sqrt{2\left((s+1)(t+1)+H_{s+1}\right)}\right)\\
&\leq \sum_{s=0}^{S}2\exp\left\{-(s+1)(t+1)\right\}\\
&\leq 2\exp\{-t\}, \qquad (2.75)
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{s=0}^{S}2\exp\left\{-(s+1)(t+1)\right\}$ can be viewed as a geometric series. Finally, notice that

$$\begin{aligned}
\sum_{s=0}^{S}2^{-s}\delta\sqrt{2\left((s+1)(t+1)+H_{s+1}\right)} &\leq \sum_{s=0}^{S}2^{-s}\delta\left(\sqrt{2(s+1)(t+1)}+\sqrt{2H_{s+1}}\right)\\
&\leq 4\delta\sqrt{1+t}+2J(\delta),
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{s=0}^{S}2^{-s}\sqrt{2(1+s)} \leq 4$ and from the inequality

$$\begin{aligned}
\sum_{s=0}^{S}2^{-s}\delta\sqrt{2H_{s+1}} &= \sum_{s=0}^{S}2^{-s}\delta\sqrt{2H_2\left(2^{-(s+1)}\delta, \mathscr{F}(\delta), \mathbb{P}_n\right)}\\
&\leq 2\sum_{s=0}^{S}\int_{2^{-(s+2)}\delta}^{2^{-(s+1)}\delta}\sqrt{2H_2\left(u, \mathscr{F}(\delta), \mathbb{P}_n\right)}\,du\\
&\leq 2J(\delta),
\end{aligned}$$

which in turn follows from the fact that $H_2\left(u, \mathscr{F}(\delta), \mathbb{P}_n\right)$ is decreasing in u. Simplifying the telescopic sum in Equation (2.75) and using the latter bounds yields that

$$P\left(\sup_{f\in\mathscr{F}(\delta)}\left[\frac{1}{\sqrt{n}}\varepsilon^\top\left(f^{S+1}-f^\star\right)\right] \geq 2J(\delta)+4\delta\sqrt{1+t}\right) \leq 2e^{-t}.$$

It follows from Equation (2.73) that

$$\begin{aligned}
&P\left(\sup_{f\in\mathscr{F}(\delta)}\left[\frac{1}{\sqrt{n}}\varepsilon^\top\left(f-f^\star\right)\right] > 2J(\delta)+4\delta\sqrt{1+t}+2^{-(S+1)}\delta\sqrt{n}\left\|\varepsilon\right\|_n\right)\\
&\qquad\leq 2e^{-t}+P\left(\sup_{f\in\mathscr{F}(\delta)}\left[\frac{1}{\sqrt{n}}\varepsilon^\top\left(f-f^{S+1}\right)\right] > 2^{-(S+1)}\delta\sqrt{n}\left\|\varepsilon\right\|_n\right).
\end{aligned}$$

However, for all $f \in \mathscr{F}(\delta)$ it holds that

$$\frac{1}{\sqrt{n}}\varepsilon^\top \left(f - f^{S+1}\right) \leq \sqrt{n}\,\|\varepsilon\|_n \left\|f - f^{S+1}\right\|_n \leq 2^{-(S+1)}\delta\sqrt{n}\,\|\varepsilon\|_n,$$

so the last term in the above bound is equal to zero. It follows that

$$\sup_{f\in\mathscr{F}(\delta)} \left[\frac{1}{\sqrt{n}}\varepsilon^\top (f - f^\star)\right] \leq 2J(\delta) + 4\delta\sqrt{1+t} + 2^{-(S+1)}\delta\sqrt{n}\,\|\varepsilon\|_n$$

with probability at least $1 - 2e^{-t}$. Notice that the LHS does not depend on S. Hence, taking the limit as $S \to \infty$ yields that

$$\sup_{f\in\mathscr{F}(\delta)} \left[\frac{1}{\sqrt{n}}\varepsilon^\top (f - f^\star)\right] \leq 2J(\delta) + 4\delta\sqrt{1+t},$$

which finishes the proof. □

The proof of the above lemma is indeed quite technical, but it illustrates the power of chaining. Most importantly, it can provide us with convergence rates for (non-parametric) least squares estimators. This is possible through the following result, which shows the connection between the entropy integral and the $\|\cdot\|_n$-distance between $\widehat{f}_n$ and $f^\star$.

Theorem 2.8.4 (Entropy integrability condition)**.** Suppose that $J(\delta) < \infty$ for all $\delta > 0$, and that $J(\delta)/\delta^2$ is a decreasing function of δ. Then, for all $t \geq 0$, and for all sequences $\{\delta_n\}_{n=1}^\infty$ of real numbers such that

$$\delta_n^2 \geq 8\left(\frac{2J(\delta_n)}{\sqrt{n}} + 4\delta_n\sqrt{\frac{1+t}{n}}\right), \tag{2.76}$$

it holds that $\left\|\widehat{f}_n - f^\star\right\|_n \leq \delta_n$ with probability at least $1 - \frac{e}{e-1}\cdot e^{-t}$.

Proof. We use the *peeling device*, namely the bound

$$P\left(\left\|\widehat{f}_n - f^\star\right\|_n > \delta_n\right) \leq \sum_{j=1}^\infty P\left(\sup_{f\in\mathscr{F}\left(2^j\delta_n\right)} \left[\frac{2}{n}\varepsilon^\top (f - f^\star)\right] \geq \left(2^{j-1}\delta_n\right)^2\right). \tag{2.77}$$

This bound holds because of the basic inequality in Lemma 2.8.2. Indeed, if $\left\|\widehat{f}_n - f^\star\right\|_n > \delta_n$ then there exists an integer $j \geq 1$ such that $\widehat{f}_n \in \mathscr{F}\left(2^j\delta_n\right) \setminus \mathscr{F}\left(2^{j-1}\delta_n\right)$. The basic inequality yields that $\frac{2}{n}\varepsilon^\top\left(\widehat{f}_n - f^\star\right) > \left(2^{j-1}\delta_n\right)^2$. Therefore

$$\sup_{f\in\mathscr{F}\left(2^j\delta_n\right)} \left[\frac{2}{n}\varepsilon^\top (f - f^\star)\right] \geq \frac{2}{n}\varepsilon^\top\left(\widehat{f}_n - f^\star\right) > \left(2^{j-1}\delta_n\right)^2.$$

We can obviously assume that this Because of the assumptions, the function

$$j \mapsto \frac{2J\left(2^j\delta_n\right) + 4\cdot 2^j\delta_n\sqrt{1+t+j}}{\left(2^j\delta_n\right)^2}$$

is decreasing, as the sum of two decreasing functions. Thus, for all $j \geq 0$,

$$\frac{2J\left(2^j\delta_n\right) + 4\cdot 2^j\delta_n\sqrt{1+t+j}}{\left(2^j\delta_n\right)^2} \leq \frac{2J\left(\delta_n\right) + 4\cdot \delta_n\sqrt{1+t}}{\delta_n^2} \leq \frac{\sqrt{n}}{8}.$$

Thus,

$$\frac{1}{2}\left(2^{j-1}\delta_n\right)^2 = \frac{1}{8}\left(2^j\delta_n\right)^2 \leq \frac{2J\left(2^j\delta_n\right)}{\sqrt{n}} + 4\cdot 2^j\delta_n\sqrt{\frac{1+t+j}{n}}$$

It follows from Equation (2.77) and from Lemma 2.8.3 that

$$\begin{aligned}
P\left(\left\|\widehat{f}_n - f^\star\right\|_n > \delta_n\right) &\leq \sum_{j=1}^{\infty} P\left(\sup_{f\in\mathscr{F}(\delta_n)}\left[\frac{2}{n}\varepsilon^\top\left(f-f^\star\right)\right] \geq \left(2^{j-1}\delta_n\right)^2\right) \\
&\leq \sum_{j=1}^{\infty} P\left(\sup_{f\in\mathscr{F}\left(2^j\delta_n\right)}\left[\frac{1}{n}\varepsilon^\top\left(f-f^\star\right)\right] \geq \frac{2J\left(2^j\delta_n\right)}{\sqrt{n}} + 4\cdot 2^j\delta_n\sqrt{\frac{1+t+j}{n}}\right) \\
&\leq 2\sum_{j=1}^{\infty}\exp\left\{-(t+j)\right\} \\
&= \frac{e}{e-1}e^{-t}.
\end{aligned}$$

□

Lastly, we present some applications of this theorem in parametric and non-parametric regression.

Example 2.8.5. Suppose that $\mathcal{Z} = \mathbb{R}^r \times \mathbb{R}$ and that $Y = f_{\theta^\star}(X) + \xi$, where $f_{\theta^\star}$ belongs to the class

$$\mathscr{F}_{\text{linear}} := \left\{f_\theta(x) = \theta_1 x_1 + \ldots + \theta_r x_r \mid \theta_1, \ldots, \theta_r \in \mathbb{R}\right\}$$

and the noise variable ξ satisfies the standard assumptions. The class $\mathscr{F}_{\text{linear}}$ is parametrized by $\theta = (\theta_1, \ldots, \theta_r) \in \mathbb{R}^r$, so ERM boils down to finding the value of θ that minimizes the empirical risk. The least squares estimator in that case is given by $\widehat{\theta}_n = X\left(X^\top X\right)^{-1}X^\top Y$, where

$$X := \begin{pmatrix} x_{11} & \ldots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nr} \end{pmatrix}$$

is the *covariate matrix* and $Y = (Y_1, \ldots, Y_n)^\top$ is the vector of the response variables. It is a standard result that

$$n \cdot \left\| f_{\widehat{\theta}_n} - f_{\theta^\star} \right\|_n^2 = \xi^\top X \left(X^\top X \right)^{-1} X^\top \xi,$$

which follows a χ_r^2 distribution. The expectation of that distribution is equal to r, so

$$\mathbb{E} \left\| f_{\widehat{\theta}_n} - f_{\theta^\star} \right\|_n^2 = \frac{r}{n}.$$

Thus, $\left\| f_{\widehat{\theta}_n} - f_{\theta^\star} \right\|_n$ converges to zero with a rate equivalent to $\sqrt{r/n}$ [16]. Let us try to derive the same convergence rate through Theorem 2.8.4. To do this, we need to check if the integrability condition $J(\delta) < \infty$ is satisfied.

We first determine a bound for the entropy of the Euclidean ball $B_r(\theta^\star, \delta)$. Let M be the packing number of this ball, and let $\{s_1, \ldots, s_M\} \subset B_r(\theta^\star, \delta)$ be a u-packing set. Then, the balls

$$B_r(s_1, u/2), \ldots, B_r(s_M, u/2)$$

are disjoint and they are all contained in the larger ball $B_r(\theta^\star, \delta + u/2)$. Comparing the total volume of the smaller balls with the volume of the larger one, we obtain

$$MC_r \cdot \left(\frac{u}{2} \right)^r \leq C_r \left(\delta + \frac{u}{2} \right)^r,$$

where C_r is the volume of the unit ball in $\mathbb{R}^r$. It follows that

$$M \left(\frac{2\delta + u}{u} \right)^r.$$

Lemma 2.5.3 yields that

$$H_2\left(u, B_r(\theta^\star, \delta), \|\cdot\|_2\right) \leq r \log \left(\frac{2\delta + u}{u} \right).$$

The mapping $\theta \mapsto \|f_\theta\|_n = \frac{1}{\sqrt{n}} \|X\theta\|_2$ defines a norm on the space range(X) and the set $\mathscr{F}_{\text{linear}}(\delta)$ is isomorphic to a δ-ball in the space range(X). Therefore, we can use the same volume argument to show that

$$H_2\left(u, \mathscr{F}_{\text{linear}}(\delta), \|\cdot\|_n\right) \leq r \log \left(\frac{2\delta + u}{u} \right).$$

[16] We keep r in the notation because in several real-world examples, r also varies with n. Such problem settings are investigated by *high-dimensional statistics*.

Hence,

$$J(\delta) \leq 2 \int_0^{\delta} \sqrt{2r} \sqrt{\log\left(1 + \frac{2\delta}{u}\right)}\, du.$$

A simple computation shows that the last integral is equal to $A\sqrt{r}\delta$, where A is a universal constant. Thus, the condition 2.76 from Theorem 2.8.4 can be written as

$$\delta_n^2 \geq 8\left(A\delta_n\sqrt{\frac{r}{n}} + 4\delta_n\sqrt{\frac{1+t}{n}}\right) \Leftrightarrow \delta_n \geq 8\left(A\sqrt{\frac{r}{n}} + 4\sqrt{\frac{1+t}{n}}\right),$$

which recovers the convergence rate $\sqrt{r/n}$ that we know from classical statistics.

Example 2.8.6 (Functions of Bounded Variation)**.** Suppose that $\mathcal{Z} = \mathbb{R} \times \mathbb{R}$ and that $x_1 \leq \ldots \leq x_n$. Consider the function class

$$\mathscr{F}_{\mathrm{BV}} := \{f : \mathbb{R} \to \mathbb{R} \mid \mathrm{TV}(f) \leq 1\}, \tag{2.78}$$

where $\mathrm{TV}(f) := \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)|$. [Birman and Solomjak, 1967] showed that, for all sufficiently small $u, \delta > 0$, it holds that $H_2(u, \mathscr{F}(\delta), \|\cdot\|_n) \leq Au^{-1}$, where $A > 0$ is a universal constant. This yields that

$$J(\delta) \leq A_0\sqrt{\delta},$$

where A_0 is another universal constant. Solving

$$\delta_n^2 \geq 8\left(\frac{2J(\delta_n)}{\sqrt{n}} + 4\delta_n\sqrt{\frac{1+t}{n}}\right),$$

we easily obtain that δ_n needs to be at least of the order of $n^{-1/3}$. This choice of δ_n satisfies the above inequality for sufficiently large $n \in \mathbb{N}$, regardless of the value of t. Therefore, $\left\|\widehat{f}_n - f^\star\right\|_n = \mathcal{O}_P\left(n^{-1/3}\right)$.

Example 2.8.7 (Lipschitz functions)**.** Suppose that $\mathcal{Z} = [0,1] \times \mathbb{R}$, and consider the function class

$$\mathscr{F}_{\mathrm{Lip}} := \{f : [0,1] \to [0,1] \mid f \text{ Lipschitz with Lipschitz constant } L = 1\}.$$

We have showed in Example 2.4.7 that $H_\infty(u, \mathscr{F}) \leq A/u$ for all $u > 0$, where A is a universal constant. Since

$$N_2(u, \mathscr{F}(\delta), \|\cdot\|_n) \leq N_2(u, \mathscr{F}, \|\cdot\|_n) \leq N_\infty(u, \mathscr{F})$$

it follows from the same argument as in the previous example that $\left\|\widehat{f}_n - f^\star\right\|_n = \mathcal{O}_P\left(n^{-1/3}\right)$.

The list of least-squares estimators for which we can derive convergence rates with respect to $\|\cdot\|_n$ using the entropy integrability condition (Theorem 2.8.4) is very large and it contains concave regression, isotonic regression, m^{th}-order Sobolev spaces, classification using indicators of convex sets and many others. The derivation of the entropy bounds is usually approached with analytic methods, like in [Birman and Solomjak, 1967]. Nevertheless, the entropy integrability condition is a very powerful and interesting method which reveals interesting and deep connections between the complexity of function classes and Glivenko-Cantelli properties.

Bibliography

[Andersen and Dobric, 1987] Andersen, N. T. and Dobric, V. (1987). The central limit theorem for stochastic processes. *The Annals of Probability*, 15(1):164–177.

[Azuma, 1967] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357 – 367.

[Billingsley, 1968] Billingsley, P. (1968). Convergence of probability measures john wiley & sons. *INC, New York*, 2(2.4).

[Billingsley, 1971] Billingsley, P. (1971). *Weak convergence of measures: Applications in probability*. SIAM.

[Billingsley, 2008] Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

[Birman and Solomjak, 1967] Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes. *Mathematics of the USSR-Sbornik*, 2(3):295.

[Cantelli, 1933] Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4(421-424).

[Chibisov, 1965] Chibisov, D. (1965). An investigation of the asymptotic power of the tests of fit. *Theory of Probability & Its Applications*, 10(3):421–437.

[Donsker, 1951] Donsker, M. D. (1951). An invariance principle for certain probability linit theorems. AMS.

[Donsker, 1952] Donsker, M. D. (1952). Justification and extension of doob's heuristic approach to the kolmogorov-smirnov theorems. *The Annals of mathematical statistics*, pages 277–281.

[Doob, 1949] Doob, J. L. (1949). Heuristic approach to the kolmogorov-smirnov theorems. *The Annals of Mathematical Statistics*, pages 393–403.

[Dudley, 1966] Dudley, R. M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois Journal of Mathematics*, 10(1):109–126.

[Dudley, 1967] Dudley, R. M. (1967). Measures on non-separable metric spaces. *Illinois Journal of Mathematics*, 11(3):449–453.

[Dudley, 1978] Dudley, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929.

[Dudley, 1984] Dudley, R. M. (1984). A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer.

[Dudley, 1985] Dudley, R. M. (1985). An extended wichura theorem, definitions of donsker class, and weighted empirical distributions. In *Probability in Banach Spaces V*, pages 141–178. Springer.

[Dudley, 1989] Dudley, R. M. (1989). *Real analysis and probability*. Cambridge University Press.

[Durrett, 2019] Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition.

[Folland, 1999] Folland, G. B. (1999). *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons.

[Glivenko, 1933] Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99.

[Hoffmann-Jørgensen, 1984] Hoffmann-Jørgensen, J. (1984). Stochastic processes on polish spaces. Unpublished.

[Hoffmann-Jørgensen, 1991] Hoffmann-Jørgensen, J. (1991). Stochastic processes on polish spaces. *Aarhus, Denmark: Various Publication Series*, 39.

[Jameson, 1974] Jameson, G. J. O. (1974). *Topology and normed spaces*. John Wiley & Sons.

[Kolmogorov, 1933] Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

[Milman et al., 2001] Milman, V. D., Schechtman, G., and Gromov, M. (2001). *Asymptotic theory of finite dimensional normed spaces*. Lecture notes in mathematics (Springer-Verlag) ; 1200. Springer, Berlin, second edition. edition.

[Pollard, 1984] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer New York, NY.

[Prokhorov, 1956] Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

[Pyke and Shorack, 1968] Pyke, R. and Shorack, G. R. (1968). Weak convergence of a two-sample empirical process and a new approach to chernoff-savage theorems. *The Annals of Mathematical Statistics*, pages 755–771.

[Skorokhod, 1956] Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290.

[van de Geer, 2000] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Van Der Vaart and Wellner, 1996] Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

[Vapnik and Chervonenkis, 1971] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

[Vershynin, 2018] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.