A Gentle Introduction to Empirical Process Theory and Applications

Bodhisattva Sen

July 19, 2022

Contents

Abstract

This document provides an introduction to the theory of empirical processes. The standard references on this topic (e.g., [van der Vaart and Wellner, 1996]) usually develop all the abstract concepts in detail before they address the statistical applications. Although this is certainly the right approach to provide a rigorous treatment of the applications, I believe that this has somewhat hindered some graduate students from appreciating the usefulness of the topic. In this set of lecture notes, I try to address a few statistical applications at the end of every section and do not go into the rigorous treatment of certain topics to make the material more accessible to a broader audience. This document arose from the lecture notes that I delivered at Stanford in Spring 2017.

As most graduate students in statistics now-a-days are not necessarily exposed to the theory of weak convergence of stochastic processes (e.g., in the space $C[0, 1]$ or $D[0, 1]$) this document tries to give the reader a brief overview of this classical theory (in Section 9). I hope this will make the transition to the theory of weak convergence on abstract spaces smoother.

I would like to thank Aditya Guntuboyina for several helpful discussions and for sharing his lecture notes on this subject (indeed, the treatment of some of the topics in this document is taken from his lecture notes[1]). I am thankful to Axel Munk and Tobias Kley[2] (who discussed some of these lecture notes in one of their seminar classes) and the students[3] in their class, and to the members of the 'Empirical Processes Reading Group' coordinated by Chao Zheng[4] at University of Southampton, for pointing out numerous typos, inconsistencies, etc. in the notes. I am also thankful to Chaowen Zheng (University of York), Myoung-Jin Keay (South Dakota State), Huiyuan Wang (Peking University), Zhen Huang (Columbia University) for pointing out further typos and inconsistencies.

Many of the examples given in this document is borrowed from the following books: [Giné and Nickl, 2016], [Koltchinskii, 2011], [Pollard, 1984], [Wainwright, 2019], [van de Geer, 2000], [van der Vaart, 1998], [van der Vaart and Wellner, 1996].

[1] see `https://www.stat.berkeley.edu/~aditya/resources/FullNotes210BSpring2018.pdf`

[2] and Shayan Hundrieser, Marcel Klatt, and Thomas Staudt

[3] Tobias W. Wegel, Erik Pudelko, Jan N. Dühmert, Meggie Marschner, Antonia Seifrid, Jana Böhm, Robin Requadt, Oliver D. Gauselmann, Tobias Weber, Huaiqing Gou, Leo H. Lehmann, Michel Groppe

[4] `http://www.personal.soton.ac.uk/cz1y20/Reading_Group/ep-group.html`

1 Introduction to empirical processes

In this chapter we introduce the main object of study (i.e., empirical processes), highlight the main questions we would like to answer, give a few historically important statistical applications that motivated the development of the field, and lay down some of the broad questions that we plan to investigate in this course.

Empirical process theory began in the 1930's and 1940's with the study of the empirical distribution function and the corresponding empirical process.

If $X_1, \ldots, X_n$ are i.i.d. real-valued random variables[5] (r.v's) with cumulative distribution function (c.d.f.) F then the *empirical distribution function* (e.d.f.) $\mathbb{F}_n : \mathbb{R} \to [0,1]$ is defined as

$$\mathbb{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i), \qquad \text{for } x \in \mathbb{R}. \tag{1}$$

In other words, for each $x \in \mathbb{R}$, the quantity $n\mathbb{F}_n(x)$ simply counts the number of X_i's that are less than or equal to x. The e.d.f. is a natural *unbiased* (i.e., $\mathbb{E}[\mathbb{F}_n(x)] = F(x)$ for all $x \in \mathbb{R}$) estimator of F. The corresponding *empirical process* is

$$\mathbb{G}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x)), \qquad \text{for } x \in \mathbb{R}. \tag{2}$$

Note that both $\mathbb{F}_n$ and $\mathbb{G}_n$ are stochastic processes[6] (i.e., random functions) indexed by the real line. By the strong law of large numbers (SLLN), for every $x \in \mathbb{R}$, we can say that

$$\mathbb{F}_n(x) \stackrel{a.s.}{\to} F(x) \qquad \text{as } n \to \infty.$$

Also, by the central limit theorem (CLT), for each $x \in \mathbb{R}$, we have

$$\mathbb{G}_n(x) \stackrel{d}{\to} N\Big(0, F(x)(1-F(x))\Big) \qquad \text{as } n \to \infty.$$

Two of the basic results in empirical process theory concerning $\mathbb{F}_n$ and $\mathbb{G}_n$ are the *Glivenko-Cantelli* and *Donsker* theorems. These results generalize the above two results to processes that hold for all x simultaneously.

---

[5]We will assume that all the random variables are defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Recall the following definitions. A *σ-field* $\mathcal{A}$ (also called σ-algebra) on a set Ω is a collection of subsets of Ω that includes the empty set, is closed under complementation, and is closed under countable unions and countable intersections. The pair $(\Omega, \mathcal{A})$ is called a *measurable space*.

If $\mathcal{C}$ is an arbitrary class of subsets of S, there is a *smallest σ-field* in S containing $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$ and called the *σ-field generated* (or induced) by $\mathcal{C}$.

A *metric* (or topological) *space* S will usually be endowed with its *Borel σ-field* $\mathcal{B}(S)$ — the σ-field generated by its topology (i.e., the collection of all open sets in S). The elements of $\mathcal{B}(S)$ are called *Borel sets*.

[6]Fix a measurable space $(S, \mathcal{S})$, an index set I, and a subset $V \subset S^I$. Then a function $W : \Omega \to V$ is called an *S-valued stochastic process* on I with *paths* in V if and only if $W_t : \Omega \to S$ is $\mathcal{S}$-measurable for every $t \in I$.

Theorem 1.1 ([Glivenko, 1933], [Cantelli, 1933]).

$$\|\mathbb{F}_n - F\|_\infty = \sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| \stackrel{a.s.}{\to} 0.$$

Theorem 1.2 ([Donsker, 1952]).

$$\mathbb{G}_n \stackrel{d}{\to} \mathbb{U}(F) \qquad \text{in } D([-\infty,\infty])^{7},$$

where $\mathbb{U}$ is the standard Brownian bridge process[8] *on* $[0,1]$.

Question: Why are uniform convergence results interesting and important?

Let us motive this by outlining a typical application of Theorems 1.1. In statistical settings, a typical use of the e.d.f. is to construct estimators of various quantities associated with the population c.d.f. Many such estimation problems can be formulated in terms of a functional γ that maps any c.d.f. F to a real number $\gamma(F)$, i.e., $F \mapsto \gamma(F)$[9]. Given a set of samples distributed according to F, the *plug-in principle* suggests replacing the unknown F with the e.d.f. $\mathbb{F}_n$, thereby obtaining $\gamma(\mathbb{F}_n)$ as an estimate of $\gamma(F)$.

For any plug-in estimator $\gamma(\mathbb{F}_n)$, an important question is to understand when it is consistent, i.e., when does $\gamma(\mathbb{F}_n)$ converge to $\gamma(F)$ in probability (or almost surely)? This question can be addressed in a unified manner for many functionals by defining a notion of continuity. Given a pair of e.d.f's F and G, let us measure the distance between them using the sup-norm

$$\|G - F\|_\infty := \sup_{x\in\mathbb{R}} |G(x) - F(x)|.$$

We can then define the continuity of a functional γ with respect to this norm: more precisely, we say that the functional γ is *continuous* at F in the *sup-norm* if, for all $\epsilon > 0$, there exists

[7] The above notion of weak convergence has not been properly defined yet. $D([-\infty,\infty])$ denotes the space of *cadlag* functions on $[-\infty,\infty]$ (French acronym: "continue à droite, limited à gauche", right continuous at each point with left limit existing at each point).

Heuristically speaking, we would say that a sequence of stochastic processes $\{\mathbb{Z}_n\}$ (as elements of $D([-\infty,\infty])$) converges in distribution to a stochastic process $\mathbb{Z}$ in $D([-\infty,\infty])$ if

$$\mathbb{E}[g(\mathbb{Z}_n)] \to \mathbb{E}[g(\mathbb{Z})], \qquad \text{as } \ n \to \infty,$$

for any bounded and continuous function $g : D([-\infty,\infty]) \to \mathbb{R}$ (ignoring measurability issues).

[8] In short, $\mathbb{U}$ is a zero-mean Gaussian process on $[0,1]$ with covariance function $\mathbb{E}[\mathbb{U}(s)\mathbb{U}(t)] = s \wedge t - st,\ s,t \in [0,1]$. To be more precise, the Brownian bridge process $\mathbb{U}$ is characterized by the following three properties:

1. $\mathbb{U}(0) = \mathbb{U}(1) = 0$. For every $t \in (0,1)$, $\mathbb{U}(t)$ is a random variable.
2. For every $k \geq 1$ and $t_1,\ldots,t_k \in (0,1)$, the random vector $(\mathbb{U}(t_1),\ldots,\mathbb{U}(t_k))$ has the $N_k(0,\Sigma)$ distribution.
3. The function $t \mapsto \mathbb{U}(t)$ is (almost surely) continuous on $[0,1]$.

[9] For example, given some integrable function $g : \mathbb{R} \to \mathbb{R}$, we may be interested in the expectation functional γ_g defined via $\gamma_g(F) := \int g(x)dF(x)$.

a $\delta > 0$ such that $\|G - F\|_\infty \le \delta$ implies that $|\gamma(G) - \gamma(F)| \le \epsilon$. This notion is useful, because for any continuous functional, it reduces the consistency question for the plug-in estimator $\gamma(\mathbb{F}_n)$ to the issue of whether or not the r.v. $\|F_n - F\|_\infty$ converges to zero.

In this course we are going to substantially generalize the two results — Theorems 1.1 and 1.2. But before we start with this endeavor, let us ask ourselves why do we need generalization of such results like. The following subsection addresses this.

1.1 Notation

The need for generalizations of Theorems 1.1 and 1.2 became apparent in the 1950's and 1960's. In particular, it became apparent that when the observations take values in a more general space $\mathcal{X}$ (such as $\mathbb{R}^d$, or a Riemannian manifold, or some space of functions, etc.), then the e.d.f. is not as natural. It becomes much more natural to consider the *empirical measure* $\mathbb{P}_n$ indexed by some class of real-valued functions $\mathcal{F}$ defined on $\mathcal{X}$.

Suppose now that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. Then the *empirical measure* $\mathbb{P}_n$ is defined by

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

where δ_x denotes the Dirac measure at x. For each $n \ge 1$, $\mathbb{P}_n$ denotes the random discrete probability measure which puts mass $1/n$ at each of the n points $X_1, \ldots, X_n$. Thus, for any Borel set $A \subset \mathcal{X}$,

$$\mathbb{P}_n(A) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_A(X_i) = \frac{\#\{i \le n : X_i \in A\}}{n}.$$

For a real-valued function f on $\mathcal{X}$, we write

$$\mathbb{P}_n(f) := \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

If $\mathcal{F}$ is a collection of real-valued functions defined on $\mathcal{X}$, then $\{\mathbb{P}_n(f) : f \in \mathcal{F}\}$ is the *empirical measure* indexed by $\mathcal{F}$. Note that the empirical measure indexed by $\mathcal{F}$ is a direct generalization of the e.d.f. in (1)); by considering

$$\mathcal{F} := \{\mathbf{1}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}\}, \qquad \text{we have} \qquad \{\mathbb{P}_n(f) : f \in \mathcal{F}\} \equiv \{\mathbb{F}_n(x) : x \in \mathbb{R}\}.$$

Let us assume that[10]

$$Pf := \int f dP$$

exists for each $f \in \mathcal{F}$. The *empirical process* $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P),$$

[10] We will use the this *operator* notation for the integral of any function f with respect to P. Note that such a notation is helpful (and preferable over the expectation notation) as then we can even treat *random* (data dependent) functions.

and the collection of random variables $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ as f varies over $\mathcal{F}$ is called the *empirical process* indexed by $\mathcal{F}$. Note that the classical empirical process (in (2)) for real-valued r.v.'s can again be viewed as the special case of the general theory for which $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \{\mathbf{1}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}\}$.

The goal of empirical process theory is to study the properties of the approximation of Pf by $\mathbb{P}_n f$, *uniformly in* $\mathcal{F}$. Traditionally, we would be concerned with probability estimates of the random quantity

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \tag{3}$$

and probabilistic limit theorems for the processes

$$\{\sqrt{n}(\mathbb{P}_n - P)f : f \in \mathcal{F}\}.$$

In particular, we will find appropriate conditions to answer the following two questions (which will extend Theorems 1.1 and 1.2):

1. **Glivenko-Cantelli**: Under what conditions on $\mathcal{F}$ does $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ converge to zero almost surely (or in probability)?

   If this convergence holds, then we say that $\mathcal{F}$ is a P-*Glivenko-Cantelli* class of functions.

2. **Donsker**: Under what conditions on $\mathcal{F}$ does $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converges as a process to some limiting object as $n \to \infty$.

   If this convergence holds, then we say that $\mathcal{F}$ is a P-*Donsker* class of functions.

Our main findings reveal that the answers (to the two above questions and more) depend crucially on the *complexity*[11] or *size* of the underlying function class $\mathcal{F}$. However, the scope of empirical process theory is much beyond answering the above two questions.

In the last 20 years there has been enormous interest in understanding the *concentration*[12] properties of $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ about its mean. In particular, one may ask if we can obtain finite sample (exponential) inequalities for the difference $\|\mathbb{P}_n - P\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$ (when $\mathcal{F}$ is uniformly bounded) in terms of the class of functions $\mathcal{F}$ and the common distribution P of $X_1, X_2, \ldots, X_n$. Talagrand's inequality ([Talagrand, 1996a]) gives an affirmative answer to this question; a result that is considered to be one of the most important and powerful results in the theory of empirical processes in the last 30 years. We will cover this topic towards the end of the course (if time permits).

[11] We will consider different geometric (packing and covering numbers) and combinatorial (shattering and combinatorial dimension) notions of complexity.

[12] Often we need to show that a random quantity $g(X_1, \ldots, X_n)$ is close to its mean $\mu(g) := \mathbb{E}[g(X_1, \ldots, X_n)]$. That is, we want a result of the form $\mathbb{P}(|g(X_1, \ldots, X_n) - \mu(g)| \geq \epsilon) \leq \delta$, for suitable ϵ and δ. Such results are known as *concentration of measure*. These results are fundamental for establishing performance guarantees of many algorithms.

The following section introduces the topic of M-estimation (also known as *empirical risk minimization*), a field that naturally relies on the study of empirical processes.

1.2 M-estimation (or empirical risk minimization)

Many problems in statistics and machine learning are concerned with estimators of the form

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{P}_n[m_\theta] = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n m_\theta(X_i). \tag{4}$$

where $X, X_1, \ldots, X_n$ denote (i.i.d.) observations from P taking values in a space $\mathcal{X}$. Here Θ denotes the parameter space and, for each $\theta \in \Theta$, m_θ denotes the a real-valued (loss-) function on $\mathcal{X}$. Such a quantity $\hat{\theta}_n$ is called an *M-estimator* as it is obtained by maximizing (or minimizing) an objective function. The map

$$\theta \mapsto -\mathbb{P}_n m_\theta = -\frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$$

can be thought of as the "empirical risk" and $\hat{\theta}_n$ denotes the empirical risk minimizer over $\theta \in \Theta$. Here are some examples:

1. **Maximum likelihood estimators**: These correspond to $m_\theta(x) = \log p_\theta(x)$.

2. **Location estimators**:

   (a) **Median**: corresponds to $m_\theta(x) = |x - \theta|$.

   (b) **Mode**: may correspond to $m_\theta(x) = \mathbf{1}\{|x - \theta| \le 1\}$.

3. **Nonparametric maximum likelihood**: Suppose $X_1, \ldots, X_n$ are i.i.d. from a density θ^* on $[0, \infty)$ that is known to be nonincreasing. Then take Θ to be the collection of all non-increasing densities on $[0, \infty)$ and $m_\theta(x) = \log \theta(x)$. The corresponding M-estimator is the MLE over all non-increasing densities. It can be shown that $\hat{\theta}_n$ exists and is unique; $\hat{\theta}_n$ is usually known as the Grenander estimator.

4. **Regression estimators**: Let $\{X_i = (Z_i, Y_i)\}_{i=1}^n$ denote i.i.d. from a regression model and let
$$m_\theta(x) = m_\theta(z, y) := -(y - \theta(z))^2,$$
for a class $\theta \in \Theta$ of real-valued functions from the domain of Z[13]. This gives the usual least squares estimator over the class Θ. The choice $m_\theta(z, y) = -|y - \theta(z)|$ gives the least absolute deviation estimator over Θ.

[13]In the simplest setting we could parametrize $\theta(\cdot)$ as $\theta_\beta(z) := \beta^\top z$, for $\beta \in \mathbb{R}^d$, in which case $\Theta = \{\theta_\beta(\cdot) : \beta \in \mathbb{R}^d\}$.

In these problems, the parameter of interest is

$$\theta_0 := \arg\max_{\theta\in\Theta} P[m_\theta].$$

Perhaps the simplest general way to address this problem is to reason as follows. By the law of large numbers, we can approximate the 'risk' for a fixed parameter θ by the empirical risk which depends only on the data, i.e.,

$$P[m_\theta] \approx \mathbb{P}_n[m_\theta].$$

If $\mathbb{P}_n[m_\theta]$ and $P[m_\theta]$ are *uniformly* close, then maybe their argmax's $\hat{\theta}_n$ and θ_0 are close. The problem is now to quantify how close $\hat{\theta}_n$ is to θ_0 as a function of the number of samples n, the dimension of the parameter space Θ, the dimension of the space $\mathcal{X}$, etc. The resolution of this question leads naturally to the investigation of quantities such as the uniform deviation

$$\sup_{\theta\in\Theta} |(\mathbb{P}_n - P)[m_\theta]|.$$

The following two examples show the importance of controlling the above display in the problem of M-estimation and classification.

Example 1.3 (Consistency of M-estimator)**.** *Consider the setup of M-estimation as introduced above where we assume that Θ is a metric space with the metric $d(\cdot,\cdot)$. In this example we describe the steps to prove the consistency of the M-estimator $\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{P}_n[m_\theta]$, as defined in (4). Formally, we want to show that*

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathbb{P}} 0 \qquad \textit{where} \qquad \theta_0 := \arg\max_{\theta\in\Theta} P[m_\theta].$$

To simplify notation we define

$$\mathbb{M}_n(\theta) := \mathbb{P}_n[m_\theta] \qquad \textit{and} \qquad M(\theta) := P[m_\theta], \quad \textit{for all } \theta\in\Theta.$$

We will assume that the class of functions $\mathcal{F} := \{m_\theta(\cdot) : \theta\in\Theta\}$ is P-Glivenko Cantelli. Fix $\delta > 0$ and let

$$\psi(\delta) := M(\theta_0) - \sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\delta} M(\theta).$$

Observe that,

$$\begin{aligned} \mathbb{P}\Big(d(\hat{\theta}_n,\theta_0)\geq\delta\Big) &\leq \mathbb{P}\Big(\mathbb{M}_n(\theta_0) \leq \sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\delta} \mathbb{M}_n(\theta)\Big) \\ &\leq \mathbb{P}\left(\sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\delta} \Big\{(\mathbb{M}_n(\theta) - M(\theta)) - (\mathbb{M}_n(\theta_0) - M(\theta_0))\Big\} \geq \psi(\delta)\right). \end{aligned}$$

Empirical process results provide bounds for the above probability (some assumptions on the relation between M and the metric $d(\cdot,\cdot)$ will be needed), e.g., we may assume that θ_0 is a well-separated maximizer, i.e., for every $\delta > 0$, $\psi(\delta) > 0$.

Note that one can further bound the above probability by

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta)/2\right),$$

but this can sometimes be too loose.

Example 1.4 (Classification)**.** *Consider a pair of random objects $X \equiv (Z, Y)$ having some joint distribution where Z takes values in a space $\mathcal{Z}$ and Y takes only two values: -1 or $+1$. A classifier is a function $g : \mathcal{Z} \to \{-1, +1\}$. The error of the classifier is given by*

$$L(g) := \mathbb{P}(g(Z) \neq Y).$$

The goal of classification is to construct a classifier with small error based on n i.i.d. observations $X_1 \equiv (Z_1, Y_1), \ldots, X_n \equiv (Z_n, Y_n)$ having the same distribution as $X = (Z, Y) \sim P$. For a classifier g, its empirical error (i.e., its error on the observed sample) is given by

$$L_n(g) := \frac{1}{n}\sum_{i=1}^{n} I\{g(Z_i) \neq Y_i\} = \mathbb{P}_n[I\{g(Z) \neq Y\}].$$

A natural strategy for classification is to select a class of classifiers $\mathcal{C}$ and then to choose the classifier in $\mathcal{C}$ which has the smallest empirical error on the observed sample, i.e.,

$$\hat{g}_n := \operatorname*{argmin}_{g\in\mathcal{C}} L_n(g).$$

How good a classifier is $\hat{g}_n$, i.e., how small is its error

$$L(\hat{g}_n) = \mathbb{P}(\hat{g}_n(Z) \neq Y | X_1, \ldots, X_n) = P(\hat{g}_n(Z) \neq Y) = \int_{(z,y):\hat{g}_n(z)\neq y} dP(z, y).$$

Two questions are relevant about $L(\hat{g}_n)$:

1. *Is $L(\hat{g}_n)$ comparable to $\inf_{g\in\mathcal{C}} L(g)$, i.e., is the error of $\hat{g}_n$ comparable to the best achievable error in the class $\mathcal{C}$?*

2. *Is $L(\hat{g}_n)$ comparable to $L_n(\hat{g}_n)$, i.e., is the error of $\hat{g}_n$ comparable to its "in-sample" empirical error?*

It is quite easy to relate these two questions to the size of $\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$. Indeed, if $g^* := \operatorname{argmin}_{g\in\mathcal{C}} L(g)$, then

$$\begin{aligned} L(\hat{g}_n) &= L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + L_n(\hat{g}_n) - L(g^*) \\ &\leq L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + L_n(g^*) - L(g^*) \quad \leq \quad L(g^*) + 2\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|, \end{aligned}$$

where we have used the fact that $L_n(\hat{g}_n) \leq L_n(g^*)$ (which follows from the definition of $\hat{g}_n$). Also,

$$L(\hat{g}_n) = L_n(\hat{g}_n) + L(\hat{g}_n) - L_n(\hat{g}_n) \leq L_n(\hat{g}_n) + \sup_{g\in\mathcal{C}} |L_n(g) - L(g)|. \tag{5}$$

Thus the key quantity to answering the above two questions is

$$\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|.$$

It is now easy to see that the above quantity is a special case of (3) *when* $\mathcal{F}$ *is taken to be the class of all functions* $I\{g(z) \neq y\}$ *as* g *varies over* $\mathcal{C}$. *Sometimes, the two inequalities above can sometimes be quite loose. Later, we shall see sharper inequalities which utilize a technique known as "localization"; see Section 8.3.*

Closely related to M-estimators are Z-estimators, which are defined as solutions to a system of equations of the form $\sum_{i=1}^{n} m_\theta(X_i) = 0$ for $\theta \in \Theta$, an appropriate function class.

We will learn how to establish *consistency*, *rates of convergence* and the *limiting distribution* for M and Z-estimators; see [van der Vaart and Wellner, 1996, Chapters 3.1-3.4] for more details.

1.3 Why study weak convergence of stochastic processes?

We quite often write a statistic of interest as a functional on the sample paths of a stochastic process in order to break the analysis of the statistic into two parts: the study of the *continuity* properties of the (measurable[14]) functional and the study of the stochastic process as a random element[15] in a space of functions. The method has its greatest appeal when many different statistics can be written as functionals on the same process, as in the following goodness-of-fit examples.

Consider the statistical problem of goodness-of-fit[16] hypothesis testing where one observes an i.i.d. sample $X_1, \ldots, X_n$ from a distribution F on the real line and wants to test the null hypothesis

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

where F_0 is a fixed continuous d.f. For testing $H_0 : F = F_0$, Kolmogorov recommended working with the quantity

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

and rejecting H_0 when D_n is large. To calculate the p-value of this test, the null distribution (i.e., the distribution of D_n under H_0) needs to be determined.

[14] Given two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a mapping $f : S \to T$ is said to be S/T-measurable or simply *measurable* if $f^{-1}(\mathcal{T}) \subset \mathcal{S}$, i.e., if

$$f^{-1}(B) := \{s \in S : f(s) \in B\} \in \mathcal{S}, \qquad \text{for every } B \in \mathcal{T}.$$

[15] A *random element of* T is a map $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (T, \mathcal{T})$ such that X is $\Omega/\mathcal{T}$-measurable [think of $(T, \mathcal{T}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in which case X is a random variable (or a random element of $\mathbb{R}$)].

[16] Many satisfactory goodness-of-fit tests were proposed by Cramér, von Mises, Kolmogorov and Smirnov. These tests are based on various *divergences* between the hypothesized c.d.f. F_0 and the e.d.f. $\mathbb{F}_n$.

Question: What is the asymptotic distribution of D_n, under H_0?

An interesting property about the null distribution of D_n is that the null distribution is the same whenever F_0 is continuous[17]. Thus we can compute the null distribution of D_n assuming that F_0 is the c.d.f. of a uniformly distributed random variable on $[0,1]$. In other words, the null distribution of D_n is the same as that of $\sup_{t\in[0,1]}|\mathbb{U}_n(t)|$ where

$$\mathbb{U}_n(t) := \sqrt{n}\,(F_n(t)-t) \quad \text{with} \quad F_n(t) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\xi_i \le t\},\ t\in[0,1],$$

and $\xi_1,\ldots,\xi_n$ are i.i.d. Unif$(0,1)$ random variables. The function $t\mapsto \mathbb{U}_n(t)$ is called the *uniform empirical process*.

The Kolmogorov-Smirnov test for $H_0 : F = F_0$ is only one of a large class of tests that are based on some measure of distance between the e.d.f. $\mathbb{F}_n$ and F_0. Another such test is the *Cramér-von Mises statistic*:

$$W_n := n\int (\mathbb{F}_n(x)-F_0(x))^2\, dF_0(x).$$

All of these quantities have the property that their null distribution (i.e., when $F = F_0$) is the same for all continuous F_0. Thus one may assume that F_0 is the uniform distribution for computing their null distribution. And in this case, all these quantities can be written in terms of the uniform empirical process $\mathbb{U}_n$.

Initially, the asymptotic distributions of these quantities were determined on a case by case basis without a unified technique. Doob realized that it should be possible to obtain these distributions using some basic properties of the uniform empirical process.

Remark 1.1 (Study of the uniform empirical process)**.** *By the multivariate CLT, for every $k\ge 1$ and $0<t_1,\ldots,t_k<1$, the random vector $(\mathbb{U}_n(t_1),\ldots,\mathbb{U}_n(t_k))$ converges in distribution to $N_k(0,\Sigma)$ where $\Sigma(i,j) := t_i\wedge t_j - t_it_j$ (here $a\wedge b := \min(a,b)$). This limiting distribution $N_k(0,\Sigma)$ is the same as the distribution of $(\mathbb{U}(t_1),\ldots,\mathbb{U}(t_k))$ where $\mathbb{U}$ is the Brownian bridge. Doob therefore conjectured that the uniform empirical process $\{\mathbb{U}_n(t) : t\in[0,1]\}$ must converge in some sense to a Brownian Bridge $\{\mathbb{U}(t) : t\in[0,1]\}$. Hopefully, this notion of convergence will be strong enough to yield that various functionals of $\mathbb{U}_n(\cdot)$ will converge to the corresponding functionals of $\mathbb{U}(\cdot)$.*

Donsker accomplished this by first establishing a rigorous theory of convergence of stochastic processes and then proving that the uniform empirical process converges to the Brownian bridge process.

Thus, we have $\mathbb{U}_n \xrightarrow{d} \mathbb{U}$ (in some sense to be defined carefully later), and thus,

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

[17] Exercise (HW1): Prove this. Hint: you may use the quantile transformation.

by appealing to the *continuous mapping theorem*[18]. Similarly, we can obtain the asymptotic distribution of the other test statistics.

In fact, there are plenty of other examples where it is convenient to break the analysis of a statistic into two parts: the study of the continuity properties of the functional and the study of the underlying stochastic process. An important example of such a "continuous" functional is the *argmax* of a stochastic process which arises in the the study of M-estimators (to be introduced in the following subsection).

1.4 Asymptotic equicontinuity

A commonly recurring theme in statistics is that we want to prove consistency or asymptotic normality of some statistic which is not a sum of independent random variables, but can be related to some natural sum of random functions indexed by a parameter in a suitable (metric) space. The following example illustrates the basic idea.

Example 1.5. *Suppose that $X, X_1, \dots, X_n, \dots$ are i.i.d. P with c.d.f. G, having a Lebesgue density g, and $\mathbb{E}(X^2) < \infty$. Let $\mu = \mathbb{E}(X)$. Consider the absolute deviations about the sample mean,*

$$M_n := \mathbb{P}_n|X - \bar{X}_n| = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_n|,$$

as an estimate of scale. This is an average of the dependent random variables $|X_i - \bar{X}_n|$. Suppose that we want to find the almost sure (a.s.) limit and the asymptotic distribution[19] *of M_n (properly normalized).*

There are several routes available for showing that $M_n \stackrel{a.s.}{\to} M := \mathbb{E}|X - \mu|$, but the method we will develop in this section proceeds as via empirical process theory. Since $\bar{X}_n \stackrel{a.s.}{\to} \mu$, we know that for any $\delta > 0$ we have $\bar{X}_n \in [\mu - \delta, \mu + \delta]$ for all sufficiently large n almost surely. Let us define, for $\delta > 0$, the random functions

$$\mathbb{M}_n(t) = \mathbb{P}_n|X - t|, \qquad \text{for } |t - \mu| \le \delta.$$

This is just the empirical measure indexed by the collection of functions

$$\mathcal{F}_\delta := \{f_t : |t - \mu| \le \delta\}, \qquad \text{where} \quad f_t(x) := |x - t|.$$

Note that $M_n \equiv \mathbb{M}_n(\bar{X}_n)$. To show that $M_n \stackrel{a.s.}{\to} M := \mathbb{E}|X - \mu|$, we write

$$\begin{aligned} M_n - M &= \mathbb{P}_n(f_{\bar{X}_n}) - P(f_\mu) \\ &= (\mathbb{P}_n - P)(f_{\bar{X}_n}) + \big[P(f_{\bar{X}_n}) - P(f_\mu)\big] \\ &= I_n + II_n. \end{aligned}$$

[18]The continuous mapping theorem states that if random elements $Y_n \stackrel{d}{\to} Y$ and g is a continuous, then $g(Y_n) \stackrel{d}{\to} g(Y)$.

[19]This example was one of the illustrative examples considered by [Pollard, 1989].

Note that,

$$|I_n| \leq \sup_{f \in \mathcal{F}_\delta} |(\mathbb{P}_n - P)(f)| \stackrel{a.s.}{\to} 0, \tag{6}$$

if $\mathcal{F}_\delta$ is P-Glivenko-Cantelli. As we will see, this collection of functions $\mathcal{F}_\delta$ is *a VC subgraph class of functions*[20] with an integrable *envelope*[21] function, and hence empirical process theory can be used to establish the desired convergence.

The convergence of the second term in II_n is easy: by the triangle inequality

$$|II_n| = \left|P(f_{\bar{X}_n}) - P(f_\mu)\right| \leq P|\bar{X}_n - \mu| = |\bar{X}_n - \mu| \stackrel{a.s.}{\to} 0.$$

Exercise (HW1): Give an alternate direct (rigorous) proof of the above result (i.e., $M_n \stackrel{a.s.}{\to} M := \mathbb{E}|X - \mu|$).

The corresponding central limit theorem is trickier. Can we show that $\sqrt{n}(M_n - M)$ converges to a normal distribution? This may still not be unreasonable to expect. After all if $\bar{X}_n$ were replaced by μ in the definition of M_n this would be an outcome of the CLT (assuming a finite variance for the X_i's) and $\bar{X}_n$ is the natural estimate of μ. Note that

$$\begin{aligned}
\sqrt{n}(M_n - M) &= \sqrt{n}(\mathbb{P}_n f_{\bar{X}_n} - P f_\mu) \\
&= \sqrt{n}(\mathbb{P}_n - P) f_\mu + \sqrt{n}(\mathbb{P}_n f_{\bar{X}_n} - \mathbb{P}_n f_\mu) \\
&= \mathbb{G}_n f_\mu + \mathbb{G}_n(f_{\bar{X}_n} - f_\mu) + \sqrt{n}(\psi(\bar{X}_n) - \psi(\mu)) \\
&= A_n + B_n + C_n \ \ \text{(say)},
\end{aligned}$$

where $\psi(t) := P(f_t) = \mathbb{E}|X - t|$. We will argue later that B_n is *asymptotically negligible* using an *equicontinuity* argument. Let us consider $A_n + C_n$. It can be easily shown that

$$\psi(t) = \mu - 2\int_{-\infty}^{t} x g(x) dx - t + 2tG(t), \qquad \text{and} \qquad \psi'(t) = 2G(t) - 1.$$

The delta method now yields:

$$A_n + C_n = \mathbb{G}_n f_\mu + \sqrt{n}(\bar{X}_n - \mu)\psi'(\mu) + o_p(1) = \mathbb{G}_n[f_\mu(X) + X\psi'(\mu)] + o_{\mathbb{P}}(1).$$

The usual CLT now gives the limit distribution of $A_n + C_n$.

Exercise (HW1): Complete the details and derive the exact form of the limiting distribution.

Definition 1.6. *Let $\{Z_n(f) : f \in \mathcal{F}\}$ be a stochastic process indexed by a class $\mathcal{F}$ equipped with a semi-metric*[22] *$d(\cdot,\cdot)$. Call $\{Z_n\}_{n\geq 1}$ to be asymptotically (or stochastically) equicontinuous at f_0 if for each $\eta > 0$ and $\epsilon > 0$ there exists a neighborhood V of f_0 for which*[23]

[20] We will formally define VC classes of functions later. Intuitively, these classes of functions have simple combinatorial properties.

[21] An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leq F(x)$, for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$.

[22] A semi-metric has all the properties of a metric except that $d(s,t) = 0$ need not imply that $s = t$.

[23] There might be measure theoretical difficulties related to taking a supremum over an uncountable set of f values, but we shall ignore these for the time being.

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{f\in V} |Z_n(f) - Z_n(f_0)| > \eta\right) < \epsilon.$$

Exercise (HW1): Show that if (i) $\{\hat{f}_n\}_{n\geq 1}$ is a sequence of (random) elements of $\mathcal{F}$ that converges in probability to f_0, and (ii) $\{Z_n(f) : f \in \mathcal{F}\}$ is asymptotically equicontinuous at f_0, then $Z_n(\hat{f}_n) - Z_n(f_0) = o_{\mathbb{P}}(1)$. [Hint: Note that with probability tending to 1, $\hat{f}_n$ will belong to each V.]

Empirical process theory offers very efficient methods for establishing the asymptotic equicontinuity of $\mathbb{G}_n$ over a class of functions $\mathcal{F}$. The fact that $\mathcal{F}$ is a *VC class* of functions with square-integrable *envelope* function will suffice to show the desired asymptotic equicontinuity.

2 Size/complexity of a function class

Let $\mathcal{F}$ be a class of measurable real-valued functions defined on $\mathcal{X}$. Whether a given class of function $\mathcal{F}$ is "Glivenko-Cantelli" or "Donsker" depends on the *size* (or *complexity*) of the class. A finite class of square integrable functions is always Donsker, while at the other extreme the class of all square integrable, uniformly bounded functions is almost never Donsker.

2.1 Covering numbers

A relatively simple way to measure the size of any set is to use *covering numbers*. Let (Θ, d) be an arbitrary semi-metric space[24]; we will assume that $\Theta \subset \Xi$ and that $d(\cdot, \cdot)$ is defined on the space Ξ. Let $\varepsilon > 0$.

Definition 2.1 (ε-cover)**.** *A ε-cover of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_N\} \subset \Xi$[25] such that for any $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ with $d(\theta, \theta_v) \le \varepsilon$.*

Definition 2.2 (Covering number)**.** *The ε-covering number of Θ is*

$$N(\varepsilon, \Theta, d) := \inf\{N \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-cover } \theta_1, \ldots, \theta_N \text{ of } \Theta\}.$$

Equivalently, the ε-covering number $N(\varepsilon, \Theta, d)$ is the minimal number of balls $B(x; \varepsilon) := \{y \in \Theta : d(x, y) \le \varepsilon\}$ of radius ε needed to cover the set Θ.

Definition 2.3 (Metric entropy)**.** *The metric entropy of the set Θ with respect to the semi-metric d is the logarithm of its covering number: $\log N(\varepsilon, \Theta, d)$.*

Note that a semi-metric space (Θ, d) is said to be *totally bounded* if the ε-covering number is finite for every $\varepsilon > 0$. We can define a related measure of size that relates to the number of disjoint balls of radius $\varepsilon > 0$ that can be placed into the set Θ.

Definition 2.4 (ε-packing)**.** *A ε-packing of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_D\} \subseteq \Theta$ such that for all distinct $v, v' \in \{1, \ldots, D\}$, we have $d(\theta_v, \theta_{v'}) > \varepsilon$.*

Definition 2.5 (Packing number)**.** *The ε-packing number of Θ is*

$$D(\varepsilon, \Theta, d) := \sup\{D \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-packing } \theta_1, \ldots, \theta_D \text{ of } \Theta\}.$$

Equivalently, call a collection of points ε-separated if the distance between each pair of points is larger than ε. Thus, the packing number $D(\varepsilon, \Theta, d)$ is the maximum number of ε-separated points in Θ.

---

[24] By a semi-metric space (Θ, d) we mean, for any $\theta_1, \theta_2, \theta_3 \in \Theta$, we have: (i) $d(\theta_1, \theta_2) = 0 \Rightarrow \theta_1 = \theta_2$; (ii) $d(\theta_1, \theta_2) = d(\theta_2, \theta_1)$; and (iii) $d(\theta_1, \theta_3) \le d(\theta_1, \theta_2) + d(\theta_2, \theta_3)$.

[25] The elements $\{\theta_1, \ldots, \theta_N\} \subset \Xi$ need not belong to Θ themselves.

A minimal ϵ-cover and or maximal ϵ-packing do not have to be finite. In the proofs of the following results, we do not separate out the case when they are infinite (in which case there is nothing show).

Lemma 2.6. *Show that*

$$D(2\varepsilon, \Theta, d) \le N(\varepsilon, \Theta, d) \le D(\varepsilon, \Theta, d), \qquad \text{for every } \varepsilon > 0.$$

Thus, packing and covering numbers have the same scaling in the radius ε.

Proof. Let us first show the second inequality. Suppose $E = \{\theta_1, \dots, \theta_D\} \subseteq \Theta$ is a maximal packing. Then for every $\theta \in \Theta \setminus E$, there exists $1 \le i \le D$ such that $d(\theta, \theta_i) \le \varepsilon$ (for if this does not hold for θ then we can construct a bigger packing set with $\theta_{D+1} = \theta$). Hence E is automatically an ε-covering. Since $N(\varepsilon, \Theta, d)$ is the minimal size of all possible coverings, we have $D(\varepsilon, \Theta, d) \ge N(\varepsilon, \Theta, d)$.

We next prove the first inequality by contradiction. Suppose that there exists a 2ε-packing $\{\theta_1, \dots, \theta_D\}$ and an ε-covering $\{x_1, \dots, x_N\}$ such that $D \ge N + 1$. Then by pigeonhole, we must have θ_i and θ_j belonging to the same ε-ball $B(x_k, \varepsilon)$ for some $i \ne j$ and k. This means that the distance between θ_i and θ_j cannot be more than the diameter of the ball, i.e., $d(\theta_i, \theta_j) \le 2\varepsilon$, which leads to a contradiction since $d(\theta_i, \theta_j) > 2\epsilon$ for a 2ε-packing. Hence the size of any 2ε-packing is less or equal to the size of any ε-covering. $\square$

Remark 2.1. *As shown in the preceding lemma, covering and packing numbers are closely related, and we can use both in the following. Clearly, they become bigger as $\varepsilon \to 0$.*

Let $\|\cdot\|$ denote any norm on $\mathbb{R}^d$. The following result gives the (order of) covering number for any bounded set in $\mathbb{R}^d$.

Lemma 2.7. *For a bounded subset $\Theta \subset \mathbb{R}^d$ there exist constants $c < C$ depending on Θ (and $\|\cdot\|$) only such that, for $\epsilon \in (0,1)$,*

$$c\left(\frac{1}{\epsilon}\right)^d \le N(\epsilon, \Theta, \|\cdot\|) \le C\left(\frac{1}{\epsilon}\right)^d.$$

Proof. If $\theta_1, \dots, \theta_D$ are ϵ-separated points in Θ, then the balls of radius $\epsilon/2$ around the θ_i's are disjoint, and their union is contained in $\Theta' := \{\theta \in \mathbb{R}^d : \|\theta - \Theta\| \le \epsilon/2\}$. Thus, the sum $Dv_d(\epsilon/2)^d$ of the volumes of these balls, where v_d is the volume of the unit ball, is bounded by $\mathrm{Vol}(\Theta')$, the volume of Θ'. This gives the upper bound of the lemma, as

$$N(\epsilon, \Theta, \|\cdot\|) \le D(\epsilon, \Theta, \|\cdot\|) \le \frac{2^d\,\mathrm{Vol}(\Theta')}{v_d}\left(\frac{1}{\epsilon}\right)^d.$$

Let $\theta_1, \dots, \theta_N$ be an ϵ-cover of Θ, i.e., the union of the balls of radius ϵ around them covers Θ. Thus the volume of Θ is bounded above by the sum of the volumes of the N

balls, i.e., by $Nv_d\epsilon^d$. This yields the lower bound of the lemma, as

$$N(\epsilon, \Theta, \|\cdot\|) \geq \frac{\mathrm{Vol}(\Theta)}{v_d}\left(\frac{1}{\epsilon}\right)^d.$$

$\square$

The following result gives an upper bound (which also happens to be optimal) on the entropy numbers of the class of Lipschitz functions[26].

Lemma 2.8. *Let $\mathcal{F} := \{f : [0,1] \to [0,1] \mid f$ is 1-Lipschitz$\}$. Then for some constant A, we have*

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq A\frac{1}{\epsilon}, \qquad \textit{for all } \epsilon > 0.$$

Proof. If $\epsilon > 1$, there is nothing to prove as then $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) = 1$ (take the function $f_0 \equiv 0$ and observe that for any $f \in \mathcal{F}$, $\|f - f_0\|_\infty \leq 1 < \epsilon$).

Let $0 < \epsilon < 1$. We will explicitly exhibit an ϵ-cover of $\mathcal{F}$ (under $\|\cdot\|_\infty$-metric) with cardinality less than $\exp(A/\epsilon)$, for some $A > 0$. This will complete the proof as $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ will then be automatically less than $\exp(A/\epsilon)$.

Let us define a ϵ-grid of the interval [0,1], i.e., $0 = a_0 < a_1 < \ldots < a_N = 1$ where $a_k := k\epsilon$, for $k = 1, \ldots, N-1$; here $N \leq \lfloor 1/\epsilon \rfloor + 1$ (where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x). Let $B_1 := [a_0, a_1]$ and $B_k := (a_{k-1}, a_k]$, $k = 2, \ldots, N$. For each $f \in \mathcal{F}$ define $\tilde{f} : [0,1] \to \mathbb{R}$ as

$$\tilde{f}(x) = \sum_{k=1}^{N} \epsilon \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \mathbf{1}_{B_k}(x). \tag{7}$$

Thus, $\tilde{f}$ is constant on the interval B_k and can only take values of the form $i\epsilon$, for $i = 0, \ldots, \lfloor 1/\epsilon \rfloor$. Observe that for $x \in B_k$ (for some $k \in \{1, \ldots, N\}$) we have

$$|f(x) - \tilde{f}(x)| \leq |f(x) - f(a_k)| + |f(a_k) - \tilde{f}(a_k)| \leq 2\epsilon,$$

where the first ϵ comes from the fact that f is 1-Lipschitz, and the second appears because of the approximation error in (7)[27]. Thus, $\|f - \tilde{f}\|_\infty \leq 2\epsilon$.

New, let us count the number of distinct $\tilde{f}$'s obtained as f varies over $\mathcal{F}$. There are at most $\lfloor 1/\epsilon \rfloor + 1$ choices for $\tilde{f}(a_1)$. Further, note that for any $\tilde{f}$ (and any $k = 2, \ldots, N$),

$$|\tilde{f}(a_k) - \tilde{f}(a_{k-1})| \leq |\tilde{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \tilde{f}(a_{k-1})| \leq 3\epsilon.$$

Therefore once a choice is made for $\tilde{f}(a_{k-1})$ there are at most 7 choices left for the next value of $\tilde{f}(a_k)$, $k = 2, \ldots, N$.

[26] Note that $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathcal{X}$.

[27] Note that, for $x \in B_k$, $\tilde{f}(x) = \tilde{f}(a_k) = \epsilon \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \leq f(a_k)$, and $f(a_k) - \tilde{f}(a_k) = \epsilon\left(\frac{f(a_k)}{\epsilon} - \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor\right) \leq \epsilon$.

Now consider the collection $\{\tilde{f} : f \in \mathcal{F}\}$. We see that this collection is a 2ϵ-cover of $\mathcal{F}$ and the number of distinct functions in this collection is upper bounded by

$$\left(\left\lfloor \frac{1}{\epsilon} \right\rfloor + 1\right) 7^{\lfloor 1/\epsilon \rfloor}.$$

Thus, $N(2\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ is bounded by the right-side of the above display, which completes the proof the result. $\square$

Thus, the set of Lipschitz functions is much "larger" than a bounded set in $\mathbb{R}^d$, since its metric entropy grows as $1/\epsilon$ as $\epsilon \to 0$, as compared to $\log(1/\epsilon)$ (cf. Lemma 2.7).

Exercise (HW1): For $L > 0$, let $\mathcal{F}_L := \{f : [0,1] \to \mathbb{R} \mid f \text{ is } L\text{-Lipschitz}\}$. Show that, for $\epsilon > 0$, $\log N(\epsilon, \mathcal{F}_L, \|\cdot\|_\infty) \geq a\frac{L}{\epsilon}$, for some constant $a > 0$. Then, using Lemma 2.8 show that $\log N(\epsilon, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\epsilon}$, for $\epsilon > 0$ sufficiently small.

2.2 Bracketing numbers

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$. We are mostly thinking of $L_r(Q)$-spaces for probability measures Q. We shall write $N(\varepsilon, \mathcal{F}, L_r(Q))$ for covering numbers relative to the $L_r(Q)$-norm $\|f\|_{Q,r} = \left(\int |f|^r dQ\right)^{1/r}$.

Definition 2.9 (ε-bracket)**.** *Given two functions $l(\cdot)$ and $u(\cdot)$, the bracket $[l, u]$ is the set of all functions $f \in \mathcal{F}$ with $l(x) \leq f(x) \leq u(x)$, for all $x \in \mathcal{X}$. An ε-bracket is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$.*

Definition 2.10 (Bracketing numbers)**.** *The bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.*

Definition 2.11 (Entropy with bracketing)**.** *The entropy with bracketing is the logarithm of the bracketing number.*

In the definition of the bracketing number, the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.

Example 2.12. *(Distribution function). When $\mathcal{F}$ is equal to the collection of all indicator functions of the form $f_t(\cdot) = \mathbf{1}_{(-\infty,t]}(\cdot)$, with t ranging over $\mathbb{R}$, then the empirical process $\mathbb{G}_n(f_t)$ is the classical empirical process $\sqrt{n}(\mathbb{F}_n(t) - F(t))$ (here $X_1, \ldots, X_n$ are i.i.d. P with c.d.f. F).*

Consider brackets of the form $[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i)}]$ for a grid points $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with the property $F(t_i-) - F(t_{i-1}) < \varepsilon$ for each $i = 1, \ldots, k$; here we assume that $\varepsilon < 1$. These brackets have $L_1(P)$-size ε. Their total number k can be chosen smaller than $2/\varepsilon$. Since $Pf^2 \leq Pf$ for every $0 \leq f \leq 1$, the $L_2(P)$-size of the brackets is bounded by $\sqrt{\varepsilon}$. Thus $N_{[\,]}(\sqrt{\varepsilon}, \mathcal{F}, L_2(P)) \leq 2/\varepsilon$, whence the bracketing numbers are of the polynomial order $1/\varepsilon^2$.

Exercise (HW1): Show that $N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N_{[\,]}(2\varepsilon, \mathcal{F}, \|\cdot\|)$, for every $\varepsilon > 0$.

In general, there is no converse inequality. Thus, apart from the constant $1/2$, bracketing numbers are bigger than covering numbers. The advantage of a bracket is that it gives pointwise control over a function: $l(x) \leq f(x) \leq u(x)$, for every $x \in \mathcal{X}$. In comparison an $L_r(P)$-ball gives integrated, but not pointwise control.

Definition 2.13 (Envelope function)**.** *An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leq F(x)$, for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. The minimal envelope function is $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$.*

Consider a class of functions $\{m_\theta : \theta \in \Theta\}$ indexed by a parameter θ in an arbitrary index set Θ with a metric d. Suppose that the dependence on θ is Lipschitz in the sense that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq d(\theta_1, \theta_2)F(x),$$

for some function $F : \mathcal{X} \to \mathbb{R}$, for every $\theta_1, \theta_2 \in \Theta$, and every $x \in \mathcal{X}$. The bracketing numbers of this class are bounded by the covering numbers of Θ as shown below.

Lemma 2.14. *Let $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ be a class of functions satisfying the preceding display for every θ_1 and θ_2 and some fixed function F. Then, for any norm $\|\cdot\|$,*

$$N_{[\,]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, \Theta, d).$$

Proof. Let $\theta_1, \ldots, \theta_p$ be an ϵ-cover of Θ (under the metric d). Then the brackets $[m_{\theta_i} - \epsilon F, m_{\theta_i} + \epsilon F]$, $i = 1, \ldots, p$, cover $\mathcal{F}$. The brackets are of size $2\epsilon\|F\|$. $\square$

Exercise (HW1): Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable function. Then for any probability measure Q and any $1 \leq r \leq \infty$,

(i) $N_{[\,]}(2\epsilon, \mathcal{F} + \mathcal{G}, L_r(Q)) \leq N_{[\,]}(\epsilon, \mathcal{F}, L_r(Q))\ N_{[\,]}(\epsilon, \mathcal{G}, L_r(Q))$;

(ii) provided $\mathcal{F}$ and $\mathcal{G}$ are bounded by 1,

$$N_{[\,]}(2\epsilon, \mathcal{F} \cdot \mathcal{G}, L_r(Q)) \leq N_{[\,]}(\epsilon, \mathcal{F}, L_r(Q))\ N_{[\,]}(\epsilon, \mathcal{G}, L_r(Q)).$$

Here, by $\mathcal{F} + \mathcal{G}$ we mean the class of functions $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$; similarly, $\mathcal{F} \cdot \mathcal{G} := \{f \cdot g : f \in \mathcal{F}, g \in \mathcal{G}\}$.

3 Glivenko-Cantelli (GC) classes of functions

Suppose that $X_1, \ldots, X_n$ are independent random variables defined on the space $\mathcal{X}$ with probability measure P. Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. The main object of study in this section is to obtain probability estimates of the random quantity

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf|.$$

The law of large numbers says that $\mathbb{P}_n f \to Pf$ almost surely, as soon as the expectation Pf exists. A class of functions is called *Glivenko-Cantelli* if this convergence is uniform in the functions belonging to the class.

Definition 3.1. *A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $P|f| < \infty$ for every $f \in \mathcal{F}$ is called Glivenko-Cantelli*[28] *(GC) if*

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to 0, \qquad \textit{almost surely.}$$

Remark 3.1 (On measurability)**.** *Note that if $\mathcal{F}$ is uncountable, $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ is the supremum of an uncountable family of random variables. In general, the supremum of uncountably many measurable functions is not necessarily measurable*[29]*. However, there are many situations when this is actually a countable supremum, e.g., in the case of the empirical distribution function (because of right continuity and existence of left limits, $\|\mathbb{F}_n - F\|_\infty = \sup_{x \in \mathbb{Q}} |\mathbb{F}_n(x) - F(x)|$, where $\mathbb{Q}$ is the set of rational numbers). Thus if $\mathcal{F}$ is countable or if there exists $\mathcal{F}_0$ countable such that $\|\mathbb{P}_n - P\|_{\mathcal{F}} = \|\mathbb{P}_n - P\|_{\mathcal{F}_0}$ a.s., then the measurability problem for $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ disappears*[30]*.*

To avoid such measurability problems we assume throughout that $\mathcal{F}$ is *pointwise measurable*, i.e., $\mathcal{F}$ contains a countable subset $\mathcal{G}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m(x) \to f(x)$ for every $x \in \mathcal{X}$[31].

In this section we prove two Glivenko-Cantelli theorems. The first theorem is the simplest and is based on entropy with bracketing. Its proof relies on finite approximation

[28] As the Glivenko-Cantelli property depends on the distribution P of the observations, we also say, more precisely, P-Glivenko-Cantelli. If the convergence is in mean or probability rather than almost surely, we speak of "Glivenko-Cantelli in mean" or "in probability".

[29] Take a non-measurable set $E \subset \mathcal{X}$ and for each $e \in E$ let $\mathbf{1}_e(\cdot)$ be the indicator function of the set $\{e\}$. Then the supremum of the (uncountable) family $\{\mathbf{1}_e(\cdot) : e \in E\}$ is the indicator function $\mathbf{1}_E(\cdot)$ (of the set E), which is not measurable.

[30] In general one can take $\|\mathbb{P}_n - P\|^*_{\mathcal{F}}$, the smallest measurable function that dominates $\|\mathbb{P}_n - P\|_{\mathcal{F}}$, which exists. $\|\cdot\|^*_{\mathcal{F}}$ works essentially as a norm, and one also has $\Pr^*\{\|\cdot\|^*_{\mathcal{F}} > t\} = \Pr\{\|\cdot\|^*_{\mathcal{F}} > t\}$, where $\Pr^*$ is *outer probability*, the infimum of the probabilities of measurable sets that contain $\mathbf{1}\{\|\cdot\|^*_{\mathcal{F}} > t\}$, and the same holds with $\mathbb{E}^*$. The calculus with $\Pr^*$ is quite similar to the usual measure theory, but there are differences (no full Fubini); see e.g., [van der Vaart and Wellner, 1996, Section 1.2].

[31] Some examples of this situation are the collection of indicators of cells in Euclidean space, the collection of indicators of balls, and collections of functions that are separable for the supremum norm.

and the law of large numbers for real variables. The second theorem uses random L_1-entropy numbers and is proved through *symmetrization* followed by a *maximal inequality.*

3.1 GC by bracketing

Theorem 3.2. *Let $\mathcal{F}$ be a class of measurable functions such that $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is Glivenko-Cantelli.*

Proof. Fix $\epsilon > 0$. Choose finitely many ϵ-brackets $[l_i, u_i]$ whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) < \epsilon$, for every i. Then, for every $f \in \mathcal{F}$, there is a bracket such that

$$(\mathbb{P}_n - P)f \le (\mathbb{P}_n - P)u_i + P(u_i - f) \le (\mathbb{P}_n - P)u_i + \epsilon.$$

Consequently,

$$\sup_{f\in\mathcal{F}}(\mathbb{P}_n - P)f \le \max_i(\mathbb{P}_n - P)u_i + \epsilon.$$

The right side converges almost surely to ϵ by the strong law of large numbers for real variables. A similar argument also yields

$$\begin{aligned} &(\mathbb{P}_n - P)f \ge (\mathbb{P}_n - P)l_i + P(l_i - f) \ge (\mathbb{P}_n - P)l_i - \epsilon \\ \Rightarrow\ &\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f \ge \min_i(\mathbb{P}_n - P)l_i - \epsilon. \end{aligned}$$

A similar argument as above (by SLLN) shows that $\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f$ is bounded below by $-\varepsilon$ almost surely. As,

$$\sup_{f\in\mathcal{F}}|(\mathbb{P}_n - P)f| = \max\left\{\sup_{f\in\mathcal{F}}(\mathbb{P}_n - P)f, -\inf_{f\in\mathcal{F}}(\mathbb{P}_n - P)f\right\},$$

we see that $\limsup \|\mathbb{P}_n - P\|_{\mathcal{F}} \le \epsilon$ almost surely, for every $\epsilon > 0$. Taking a sequence $\epsilon_m \downarrow 0$ yields the desired result. ☐

Example 3.3. *(Distribution function). The previous proof generalizes a well-known proof of the classical GC theorem for the e.d.f. on the real line. Indeed, the set of indicator functions of cells $(-\infty, c]$ possesses finite bracketing numbers for any underlying distribution; simply use the brackets $\left[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i)}\right]$ for a grid of points $-\infty = t_0 < t_1 < \ldots, t_k = +\infty$ with the property $P(t_{i-1}, t_i) < \epsilon$ for each i.*

Example 3.4 (Pointwise compact class)**.** *Let $\mathcal{F} = \{m_\theta(\cdot) : \theta \in \Theta\}$ be a collection of measurable functions with integrable envelope function F indexed by a compact metric space Θ such that the map $\theta \mapsto m_\theta(x)$ is continuous for every x. Then the bracketing numbers of $\mathcal{F}$ are finite and hence $\mathcal{F}$ is Glivenko-Cantelli.*

We can construct the brackets in the obvious way in the form $[m_B, m^B]$, where B is an open ball and m_B and m^B are the infimum and supremum of m_θ for $\theta \in B$, respectively (i.e., $m_B(x) = \inf_{\theta\in B} m_\theta(x)$, and $m^B(x) = \sup_{\theta\in B} m_\theta(x)$).

Given a sequence of balls B_k with common center a given θ and radii decreasing to 0, we have $m^{B_k} - m_{B_k} \downarrow m_\theta - m_\theta = 0$ by the continuity, pointwise in x, and hence also in L_1 by the dominated convergence theorem and the integrability of the envelope. Thus, given $\epsilon > 0$, for every θ there exists a ball B around θ such that the bracket $[m_B, m^B]$ has size at most ϵ. By the compactness of Θ, the collection of balls constructed in this way has a finite subcover. The corresponding brackets cover $\mathcal{F}$. This construction shows that the bracketing numbers are finite, but it gives no control on their sizes.

An example of such a class would be the log-likelihood function of a parametric model $\{p_\theta(x) : \theta \in \Theta\}$, where $\Theta \in \mathbb{R}^d$ is assumed to be compact[32] *and $p_\theta(x)$ is assumed to be continuous in θ for P_{θ_0}-a.e. x.*

The goal for the remainder of this section is to prove the following theorem.

Theorem 3.5 (GC by entropy)**.** *Let $\mathcal{F}$ be a class of measurable functions with envelope F such that $P(F) < \infty$. Let $\mathcal{F}_M$ be the class of functions $f\mathbf{1}\{F \le M\}$ where f ranges over $\mathcal{F}$. Then $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ almost surely if and only if*

$$\frac{1}{n} \log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0, \tag{8}$$

for every $\epsilon > 0$ and $M > 0$[33]*. In that case the convergence takes place in mean also.*

Both the statement and the proof of the GC theorem with entropy are more complicated than the previous bracketing theorem. However, the result gives a precise (necessary and sufficient) characterization for a class of functions to be GC. Moreover, the sufficiency condition for the GC property can be checked for many classes of functions by elegant combinatorial arguments, as will be discussed later. The proof of the above result needs numerous other concepts, which we introduce below.

3.2 Preliminaries

In this subsection we will introduce a variety of simple results that will be useful in proving the Glivenko-Cantelli theorem with entropy. We will expand on each of these topics later on, as they indeed form the foundations of empirical process theory.

3.2.1 Hoeffding's inequality for the sample mean

Bounds on the tail probability of (the maximum of a bunch of) random variables form the backbone of most of the results in empirical process theory (e.g., GC theorem, maximal

[32] This is a stringent assumption in many situations. If we assume that m_θ is convex/concave (in θ), then it suffices to consider compact subsets of the parameter space (we will see such an example soon; also see e.g., [Hjort and Pollard, 2011] for a more refined approach). In other situations we have to argue from first principles that it is enough to restrict attention to compacts.

[33] Furthermore, the random entropy condition is necessary.

inequalities needed to show asymptotic equicontinuity, etc.). In this subsection we will review some basic results in this topic which will be used to prove the GC theorem with entropy. We start with a very simple (but important and useful) result.

Lemma 3.6 (Markov's inequality)**.** *Let $Z \geq 0$ be a random variable. Then for any $t > 0$,*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t}.$$

Proof. Observe that $t\mathbf{1}\{Z \geq t\} \leq Z$, which on taking expectations yield the above result. ☐

The above lemma implies Chebyshev's inequality.

Lemma 3.7 (Chebyshev's inequality)**.** *If Z has a finite variance* $\mathrm{Var}(Z)$*, then*

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \frac{\mathrm{Var}(Z)}{t^2}.$$

This above lemma is probably the simplest *concentration* inequality. Thus, if $X_1, \ldots, X_n$ are i.i.d. with finite variance σ^2, and if $Z = \sum_{i=1}^n X_i =: S_n$, then (by independence) $\mathrm{Var}(Z) = n\sigma^2$, and

$$\mathbb{P}\left(\Big|S_n - \mathbb{E}S_n\Big| \geq t\sqrt{n}\right) \leq \frac{\sigma^2}{t^2}.$$

Thus, the typical deviations/fluctuations of the sum of n i.i.d. random variables are at most of order $\sqrt{n}$. However, by the central limit, theorem (CLT), for fixed $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(S_n - \mathbb{E}S_n \geq t\sqrt{n}\right) = 1 - \Phi\left(\frac{t}{\sigma}\right) \leq \frac{\sigma}{\sqrt{2\pi}t}\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

where the last inequality uses a standard bound on the normal CDF. Thus, we see that although Chebyshev's inequality gets the order $\sqrt{n}$ correct, the dependence on t^2/σ^2 is not as predicted by the CLT (we expect an exponential decrease in t^2/σ^2).

Indeed, we can improve the above bound by assuming that Z has a moment generating function. The main trick here is to use Markov's inequality in a clever way: if $\lambda > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z > t) = \mathbb{P}\left(e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda t}\right) \leq \frac{\mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}]}{e^{\lambda t}}. \tag{9}$$

Now, we can derive bounds for the *moment generating function* $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$ and optimize over λ.

When dealing with S_n, we can extend the idea as follows. Observe that,

$$\mathbb{E}e^{\lambda Z} = \mathbb{E}\left(\prod_{i=1}^n e^{\lambda X_i}\right) = \prod_{i=1}^n \mathbb{E}e^{\lambda X_i}, \qquad \text{(by independence)}. \tag{10}$$

Now it suffices to find bounds for $\mathbb{E}e^{\lambda X_i}$.

Lemma 3.8 (Exercise (HW1))**.** *Let X be a random variable with $\mathbb{E}X = 0$ and $X \in [a, b]$ with probability 1 (w.p.1). Then, for any $\lambda > 0$,*

$$\mathbb{E}(e^{\lambda X}) \le e^{\lambda^2 (b-a)^2/8}.$$

The above lemma, combined with (9) and (10), implies the following Hoeffding's tail inequality.

Lemma 3.9 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ w.p.1. Then, we obtain,*

$$\mathbb{P}\left(S_n - \mathbb{E}S_n \ge t\right) \le e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2},$$

and

$$\mathbb{P}\left(S_n - \mathbb{E}S_n \le -t\right) \le e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

Proof. For $\lambda, t \ge 0$, Markov's inequality and the independence of X_i implies:

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}S_n \ge t\right) &= \mathbb{P}\left(e^{\lambda(S_n - \mathbb{E}[S_n])} \ge e^{\lambda t}\right) \le e^{-\lambda t}\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \\
&= e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \le e^{-\lambda t}\prod_{i=1}^n e^{\frac{\lambda^2(b_i - a_i)^2}{8}} \\
&= \exp\Big(-\lambda t + \tfrac{1}{8}\lambda^2 \sum_{i=1}^n (b_i - a_i)^2\Big).
\end{aligned}$$

To get the best possible upper bound, we find the minimum of the right hand side of the last inequality as a function of λ. Define $g : \mathbb{R}_+ \to \mathbb{R}$ such that $g(\lambda) = -\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2$. Note that g is a quadratic function and achieves its minimum at $\lambda = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$. Plugging in this value of λ in the above bound we obtain the desired result. We can similarly prove the tail bound for $t < 0$. ☐

Example 3.10 (Hoeffding's bound for i.i.d. random variables)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. such that $X_1 \in [a, b]$ w.p. 1. Then for any given $\alpha \in (0, 1)$ a direct consequence of Lemma 3.9 show that*

$$S_n - \mathbb{E}[S_n] \ge \sqrt{\frac{n(b-a)^2}{2}\log\frac{1}{\alpha}} \tag{11}$$

with probability at most α. In fact, by Hoeffding's inequality, we can obtain an $1 - \alpha$ honest conservative confidence (symmetric) interval (around the sample mean $\bar{X}_n$) for the population mean $\mu := \mathbb{E}[X_1]$ as:

$$\left[\bar{X}_n - \frac{b-a}{\sqrt{2n}}\sqrt{\log\frac{2}{\alpha}},\ \ \bar{X}_n + \frac{b-a}{\sqrt{2n}}\sqrt{\log\frac{2}{\alpha}}\right].$$

Hoeffding's inequality does not depend on the distribution of the X_i's (which is good), but also does not incorporate the dependence on $\mathrm{Var}(X_i)$ in the bound (which can result in an inferior bound; e.g., consider $X_i \sim \text{Bernoulli}(p)$ where p is close to 0 or 1).

3.2.2 Sub-Gaussian random variables/processes

A sub-Gaussian distribution is a probability distribution with strong tail decay property. Informally, the tails of a sub-Gaussian distribution are dominated by (i.e., decay at least as fast as) the tails of a Gaussian.

Definition 3.11. *X is said to be sub-Gaussian if $\exists$ constants $C, v > 0$ s.t. $\mathbb{P}(|X| > t) \leq Ce^{-vt^2}$ for every $t > 0$.*

The following are equivalent characterizations of a sub-Gaussian random variable (Exercise: HW1)

- The distribution of X is sub-Gaussian.
- There exists $a > 0$ such that $\mathbb{E}[e^{aX^2}] < +\infty$.
- Laplace transform condition: $\exists B, b > 0$ such that $\forall\ \lambda \in \mathbb{R}$, $\mathbb{E}e^{\lambda(X - \mathbb{E}[X])} \leq Be^{\lambda^2 b}$.
- Moment condition: $\exists\ K > 0$ such that for all $p \geq 1$, $(\mathbb{E}|X|^p)^{1/p} \leq K\sqrt{p}$.

Suppose (T, d) is a semi-metric space and let $\{X_t, t \in T\}$ be a stochastic process indexed by T satisfying

$$\mathbb{P}\left(|X_s - X_t| \geq u\right) \leq 2\exp\left(-\frac{u^2}{2d(s,t)^2}\right) \qquad \text{for all } u > 0. \tag{12}$$

Such a stochastic process is called *sub-Gaussian* with respect to the semi-metric d. Any Gaussian process is sub-Gaussian for the standard deviation semi-metric $d(s,t) = \sqrt{\mathrm{Var}(X_s - X_t)}$. Another example is the *Rademacher process*

$$X_a := \sum_{i=1}^{n} a_i \varepsilon_i, \qquad a := (a_1, \ldots, a_n) \in \mathbb{R}^n,$$

for independent Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$, i.e.,

$$\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2};$$

variables that are $+1$ or -1 with probability $1/2$ each. The result follows directly from the following lemma.

Lemma 3.12 (Hoeffding's inequality for Rademacher variables)**.** *Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ be a vector of constants and $\varepsilon_1, \ldots, \varepsilon_n$ be Rademacher random variables. Then*

$$\mathbb{P}\Big(\big|\sum_{i=1}^{n} a_i \varepsilon_i\big| \geq x\Big) \leq 2e^{-x^2/(2\|a\|^2)},$$

where $\|a\|$ denotes the Euclidean norm of a.

Proof. For any λ and Rademacher variable ε, one has $\mathbb{E}e^{\lambda\varepsilon} = (e^{\lambda} + e^{-\lambda})/2 \le e^{\lambda^2/2}$, where the last inequality follows after writing out the power series. Thus, by Markov's inequality, for any $\lambda > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n} a_i\varepsilon_i \ge x\Big) \le e^{-\lambda x}\, \mathbb{E}e^{\lambda \sum_{i=1}^{n} a_i\varepsilon_i} \le e^{(\lambda^2/2)\|a\|^2 - \lambda x}.$$

The best upper bound is obtained for $\lambda = x/\|a\|^2$ and is the exponential in the probability of the lemma. Combination with a similar bound for the lower tail yields the probability bound. $\square$

Here is a useful (and probably the earliest and simplest) 'maximal inequality' which is an application of Hoeffding's inequality.

Lemma 3.13. *Suppose that $Y_1, \ldots, Y_N$ (not necessarily independent) are sub-Gaussian in the sense that, for all $\lambda > 0$,*

$$\mathbb{E}e^{\lambda Y_i} \le e^{\lambda^2\sigma^2/2}, \qquad \textit{for all } i = 1, \ldots, N.$$

Then,

$$\mathbb{E} \max_{i=1,\ldots,N} Y_i \le \sigma\sqrt{2\log N}. \tag{13}$$

Proof. Observe that

$$e^{\lambda \mathbb{E}\max_{i=1,\ldots,N} Y_i} \le \mathbb{E}e^{\lambda \max_{i=1,\ldots,N} Y_i} \le \sum_{i=1}^{N} \mathbb{E}e^{\lambda Y_i} \le N e^{\lambda^2\sigma^2/2},$$

where we have used Jensen's inequality in the first step (taking the function $e^{\lambda x}$). Taking logarithms yields

$$\mathbb{E}\left[\max_{i=1,\ldots,N} Y_i\right] \le \frac{\log N}{\lambda} + \frac{\lambda\sigma^2}{2}.$$

Optimizing with respect to λ (differentiating and then equating to 0) yields the result. $\square$

Example 3.14. *Let $A_1, \ldots, A_N \subset \mathcal{X}$ and let $X_1, \ldots, X_n$ be i.i.d. random points in $\mathcal{X}$. Let*

$$P(A) = \mathbb{P}(X_i \in A) \qquad \textit{and} \qquad \mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(X_i).$$

By Hoeffding's inequality, for each A,

$$\mathbb{E}e^{\lambda(\mathbb{P}_n(A) - P(A))} = \mathbb{E}e^{(\lambda/n)\sum_{i=1}^{n}(\mathbf{1}_A(X_i) - P(A))} = \prod_{i=1}^{n} \mathbb{E}e^{(\lambda/n)(\mathbf{1}_A(X_i) - P(A))} \le e^{\lambda^2/(8n)}.$$

Thus, by a simple maximal inequality (see Lemma 3.13),

$$\mathbb{E} \max_{i=1,\ldots,N} (\mathbb{P}_n(A) - P(A)) \le \sqrt{\frac{\log N}{2n}}.$$

Lemma 3.13 can be easily extended to yield the following maximal inequality.

Lemma 3.15. *Let ψ be a strictly increasing, convex, nonnegative function. If $\xi_1, \ldots, \xi_N$ are random variables such that*

$$\mathbb{E}[\psi(|\xi_i|/c_i)] \le L, \qquad \text{for } i = 1, \ldots, N,$$

where L is a constant, then

$$\mathbb{E} \max_{1 \le i \le N} |\xi_i| \le \psi^{-1}(LN) \max_{1 \le i \le N} c_i.$$

Proof. By the properties of ψ,

$$\psi\left(\frac{\mathbb{E}\max|\xi_i|}{\max c_i}\right) \le \psi\left(\mathbb{E}\max\frac{|\xi_i|}{c_i}\right) \le \mathbb{E}\psi\left(\max\frac{|\xi_i|}{c_i}\right) \le \sum_{i=1}^N \mathbb{E}\psi\left(\frac{|\xi_i|}{c_i}\right) \le LN.$$

Applying ψ^{-1} to both sides we get the result. □

Lemma 3.16. *Show that if ξ_i's are linear combinations of Rademacher variables (i.e., $\xi_i = \sum_{k=1}^n a_k^{(i)} \varepsilon_k$, $a^{(i)} = (a_1^{(i)}, \ldots, a_n^{(i)}) \in \mathbb{R}^n$), then: (i) $\mathbb{E}[e^{\xi_i^2/(6\|a^{(i)}\|^2)}] \le 2$; (ii) by using $\psi(x) = e^{x^2}$ show that, for $N \ge 2$,*

$$\mathbb{E} \max_{1 \le i \le N} |\xi_i| \le C\sqrt{\log N} \max_{1 \le i \le N} \|a^{(i)}\|. \tag{14}$$

(iii) Further, show that $C = 2\sqrt{6}$ for Rademacher linear combinations.

Exercise (HW1): (a) An *Orlicz function* is a convex, increasing function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ with $0 \le \psi(0) < 1$ (most authors actually require $\psi(0) = 0$).

Define the *Orlicz norm* $\|X\|_\psi$ (seminorm actually, unless one identifies random variables that are almost everywhere equal) by

$$\|X\|_\psi := \inf\{c > 0 : \mathbb{E}[\psi(|X|/c)] \le 1\},$$

with the understanding that $\|X\|_\psi = \infty$ if the infimum runs over an empty set. Let $\psi(x) = \exp(x^2) - 1$. Then show that $\|X\|_\psi < \infty$ if and only if $X - \mathbb{E}X$ is sub-Gaussian.

3.3 Symmetrization

Symmetrization (or randomization) technique plays an essential role in empirical process theory. The symmetrization replaces $\sum_{i=1}^n (f(X_i) - Pf)$ by $\sum_{i=1}^n \varepsilon_i f(X_i)$ with i.i.d. Rademacher[34] random variables $\varepsilon_1, \ldots, \varepsilon_n$ independent of $X_1, \ldots, X_n$. Note that $\sum_{i=1}^n \varepsilon_i f(X_i)$ can be thought of as the correlation between the vector $(f(X_1), \ldots, f(X_n))$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$. Thus,

$$\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|$$

[34] Recall that a Rademacher random variable ε takes values ± 1 with equal probability 1/2.

denotes the maximum correlation taken over all functions $f \in \mathcal{F}$. The intuition here is: a function class is extremely large — and, in fact, "too large" for statistical purposes — if we can always find a function (in the class) that has a high correlation with a randomly drawn noise vector.

The advantage of symmetrization lies in the fact that the symmetrized process is typically easier to control than the original process, as we will find out in several places. For example, even though $\sum_{i=1}^n (f(X_i) - Pf)$ has only low order moments, $\sum_{i=1}^n \varepsilon_i f(X_i)$ is *sub-Gaussian*, conditionally on $X_1, \ldots, X_n$. In what follows, $\mathbb{E}_\varepsilon$ denotes the expectation with respect to $\varepsilon_1, \varepsilon_2, \ldots$ only; likewise, $\mathbb{E}_X$ denotes the expectation with respect to $X_1, X_2, \ldots$ only.

The *symmetrized empirical measure* and *process* are defined by

$$f \mapsto \mathbb{P}_n^o f := \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i), \qquad f \mapsto \mathbb{G}_n^o f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i).$$

The symmetrized empirical processes has mean function zero.

One main approach to proving empirical limit theorems is to pass from $\mathbb{P}_n - P$ to $\mathbb{P}_n^o$ and next apply arguments conditionally on the original X's. The idea is that, for fixed $X_1, \ldots, X_n$, the symmetrized empirical measure is a Rademacher process, hence a sub-Gaussian process to which Dudley's entropy result (see Chapter 4) can be applied.

Theorem 3.17 (Symmetrization)**.** *For any class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \le 2\,\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}.$$

Proof. Let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$, and defined on the same probability space. For fixed values $X_1, \ldots, X_n$,

$$\begin{aligned}\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \frac{1}{n}\Big|\sum_{i=1}^n \{f(X_i) - \mathbb{E}f(Y_i)\}\Big| \quad &= \quad \sup_{f\in\mathcal{F}} \frac{1}{n}\Big|\mathbb{E}_Y\big[\sum_{i=1}^n \{f(X_i) - f(Y_i)\}\big]\Big| \\ &\le \quad \frac{1}{n}\mathbb{E}_Y\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \{f(X_i) - f(Y_i)\}\Big|\Big]\end{aligned}$$

where $\mathbb{E}_Y$ is the expectation with respect to $Y_1, \ldots, Y_n$, given fixed values of $X_1, \ldots, X_n$, and we have used the fact that for a class of functions $\mathcal{G}$, $\sup_{g\in\mathcal{G}} \mathbb{E}[g(Z)] \le \mathbb{E}\left[\sup_{g\in\mathcal{G}} |g(Z)|\right]$ for some random vector Z.

Taking the expectation with respect to $X_1, \ldots, X_n$, we get

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \mathbb{E}_X\|\mathbb{P}_n - P\|_{\mathcal{F}} \le \mathbb{E}_{X,Y}\Big\|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\Big\|_{\mathcal{F}}.$$

Adding a minus sign in front of a term $[f(X_i) - f(Y_i)]$ has the effect of exchanging X_i and Y_i. Because the Y's are independent copies of the X's, the expectation of any function

$g(X_1,\ldots,X_n,Y_1,\ldots,Y_n)$ ($=\sup_{f\in\mathcal{F}}|\sum_{i=1}^n[f(X_i)-f(Y_i)]|$ here) remains unchanged under permutations of its $2n$ arguments. Hence the expression

$$\mathbb{E}\frac{1}{n}\Big\|\sum_{i=1}^n e_i[f(X_i)-f(Y_i)]\Big\|_{\mathcal{F}}$$

is the same for any n-tuple $(e_1,\ldots,e_n)\in\{-1,+1\}^n$. Deduce that

$$\mathbb{E}\|\mathbb{P}_n-P\|_{\mathcal{F}}\le\mathbb{E}_\varepsilon\mathbb{E}_{X,Y}\Big\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i[f(X_i)-f(Y_i)]\Big\|_{\mathcal{F}}.$$

Using the triangle inequality to separate the contributions of the X's and the Y's and noting that they are both equal to $\mathbb{E}\|\mathbb{P}_n^o\|_{\mathcal{F}}$. □

Remark 3.2. *The symmetrization lemma is valid for any class $\mathcal{F}$. In the proofs of Glivenko-Cantelli and Donsker theorems, it will be applied not only to the original set of functions of interest, but also to several classes constructed from such a set $\mathcal{F}$ (such as the class $\mathcal{F}_\delta$ of small differences). The next step in these proofs is to apply a maximal inequality to the right side of the above theorem, conditionally on $X_1,\ldots,X_n$.*

Lemma 3.18 (A more general version of the symmetrization lemma; Exercise (HW1))**.** *Suppose that $Pf=0$ for all $f\in\mathcal{F}$. Let $\varepsilon_1,\ldots,\varepsilon_n$ be independent Rademacher random variables independent of $X_1,\ldots,X_n$. Let $\Phi:\mathbb{R}_+\to\mathbb{R}_+$ be a nondecreasing convex function, and let $\mu:\mathcal{F}\to\mathbb{R}$ be a bounded functional such that $\{f+\mu(f):f\in\mathcal{F}\}$ is pointwise measurable. Then,*

$$\mathbb{E}\left[\Phi\Big(\frac{1}{2}\Big\|\sum_{i=1}^n\varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\Big)\right]\le\mathbb{E}\left[\Phi\Big(\Big\|\sum_{i=1}^n f(X_i)\Big\|_{\mathcal{F}}\Big)\right]\le\mathbb{E}\left[\Phi\Big(2\Big\|\sum_{i=1}^n\varepsilon_i(f(X_i)+\mu(f))\Big\|_{\mathcal{F}}\Big)\right].$$

3.4 Proof of GC by entropy

In this subsection we prove Theorem 3.5. We only show the sufficiency of the entropy condition. By the symmetrization result, measurability of the class $\mathcal{F}$, and Fubini's theorem,

$$\begin{aligned}\mathbb{E}\|\mathbb{P}_n-P\|_{\mathcal{F}} &\le 2\mathbb{E}_X\mathbb{E}_\varepsilon\Big\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\\ &\le 2\mathbb{E}_X\mathbb{E}_\varepsilon\Big\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\Big\|_{\mathcal{F}_M}+2P[F\mathbf{1}\{F>M\}],\end{aligned}$$

by the triangle inequality, for every $M>0$. For sufficiently large M, the last term is arbitrarily small. To prove convergence in mean, it suffices to show that the first term converges to zero for fixed M. Fix $X_1,\ldots,X_n$. If $\mathcal{G}$ is an η-net[35] in $L_1(\mathbb{P}_n)$ over $\mathcal{F}_M$, then for any $f\in\mathcal{F}_M$, there exists $g\in\mathcal{G}$ such that

$$\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\Big|\le\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i g(X_i)\Big|+\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i[f(X_i)-g(X_i)]\Big|\le\Big\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i g(X_i)\Big\|_{\mathcal{G}}+\eta.$$

[35]The set of centers of balls of radius η that cover T is called an ϵ-net of T.

Thus,

$$\mathbb{E}_\varepsilon\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}_M} \le \mathbb{E}_\varepsilon\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\Big\|_{\mathcal{G}} + \eta. \tag{15}$$

The cardinality of $\mathcal{G}$ can be chosen equal to $N(\eta, \mathcal{F}_M, L_1(\mathbb{P}_n))$. Given $X_1, \ldots, X_n$, the symmetrized empirical process $\mathbb{G}_n^o$ is sub-Gaussian with respect to the $L_2(\mathbb{P}_n)$-norm. Using the maximal inequality in Lemma 3.15 with $\psi(x) = \exp(x^2)$ (see (14)) shows that the preceding display does not exceed

$$C\sqrt{\log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))}\frac{1}{\sqrt{n}}\sup_{g\in\mathcal{G}}\|g\|_n + \epsilon,$$

where $\|g\|_n^2 := \sum_{i=1}^n g(X_i)^2/n$ and C is a universal constant.

The $\|\cdot\|_n$ norms of $g \in \mathcal{G}$ are bounded above by M (this can always be ensured by truncating g if required). By assumption the square root of the entropy divided by $\sqrt{n}$ tends to zero in probability. Thus the right side of the above display tends to ϵ in probability. Since this argument is valid for every $\epsilon > 0$, it follows that the left side of (15) converges to zero in probability.

Next we show that $\mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_M}|X_1, \ldots, X_n\right]\right]$ converges to 0. Since $\mathbb{E}_\varepsilon\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}_M}$ is bounded by M and converges to 0 (in probability), its expectation with respect to $X_1, \ldots, X_n$ converges to zero by the dominated convergence theorem.

This concludes the proof that $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ in mean. That it also converges almost surely follows from the fact that the sequence $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ is a reverse sub-martingale with respect to a suitable filtration; see e.g., [van der Vaart and Wellner, 1996, Lemma 2.4.5] (by martingale theory any nonnegative reverse sub-martingale converges almost surely to some limit[36]). □

3.5 Applications

3.5.1 Consistency of M/Z-estimators

Consider the setup of M-estimation as introduced in Section 1.2 and Example 1.3 where we assume that Θ is a metric space with the metric $d(\cdot,\cdot)$. In this example we describe the steps to prove the *consistency* of the M-estimator $\hat\theta_n := \arg\max_{\theta\in\Theta}\mathbb{P}_n[m_\theta]$, as defined in (4). Formally, we want to show that

$$d(\hat\theta_n, \theta_0) \xrightarrow{\mathbb{P}} 0 \qquad \text{where} \qquad \theta_0 := \arg\max_{\theta\in\Theta} P[m_\theta].$$

To simplify notation we define

$$\mathbb{M}_n(\theta) := \mathbb{P}_n[m_\theta] \qquad \text{and} \qquad M(\theta) := P[m_\theta], \quad \text{for all } \ \theta \in \Theta.$$

[36]**Result**: If $PF < \infty$, then $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ converges almost surely and in L_1.

We will assume that the class of functions $\mathcal{F} := \{m_\theta(\cdot) : \theta \in \Theta\}$ is P-Glivenko Cantelli. We will further need to assume that θ_0 is a *well-separated* maximizer, i.e., for every $\delta > 0$,

$$M(\theta_0) > \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta).$$

Fix $\delta > 0$ and let

$$\psi(\delta) := M(\theta_0) - \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta).$$

Observe that,

$$\begin{aligned}
\{d(\hat{\theta}_n, \theta_0) \geq \delta\} \quad &\Rightarrow \quad M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \delta} M(\theta) \\
&\Leftrightarrow \quad M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\
&\Rightarrow \quad M(\hat{\theta}_n) - M(\theta_0) + (\mathbb{M}_n(\theta_0) - \mathbb{M}_n(\hat{\theta}_n)) \leq -\psi(\delta) \\
&\Rightarrow \quad 2 \sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta).
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(d(\hat{\theta}_n, \theta_0) \geq \delta\right) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\delta)/2\right) \to 0$$

by the fact that $\mathcal{F}$ is P-Glivenko Cantelli.

The above result of course assumes that $\mathcal{F}$ is P-Glivenko Cantelli, which we can verify by showing the sufficient conditions in Theorem 3.2 or Theorem 3.5 hold. For example, for a pointwise compact space, as described in Example 3.4, this immediately yields the consistency of $\hat{\theta}_n$.

Note that the sufficient conditions needed to show that $\mathcal{F}$ is P-GC is hardly ever met when Θ is not a compact set (observe that the finiteness of covering numbers necessarily imply that the underlying set is totally bounded). We give a lemma below which replaces compactness by a *convexity* assumption.

Lemma 3.19 (Exercise (HW1))**.** *Suppose that Θ is a convex subset of $\mathbb{R}^d$, and that $\theta \mapsto m_\theta(x)$, is continuous and concave, for all $x \in \mathcal{X}$. Suppose that $\mathbb{E}[G_\epsilon(X)] < \infty$ where $G_\epsilon(x) := \sup_{\|\theta - \theta_0\| \leq \epsilon} |m_\theta(x)|$, for $x \in \mathcal{X}$. Then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.*

Hint: Define $\alpha := \frac{\epsilon}{\epsilon + \|\hat{\theta}_n - \theta_0\|}$ and $\tilde{\theta}_n := \alpha \hat{\theta}_n + (1 - \alpha)\theta_0$ and compare $\mathbb{M}_n(\tilde{\theta}_n)$ with $\mathbb{M}_n(\theta_0)$.

3.5.2 Consistency of least squares regression

Suppose that we have data

$$Y_i = g_0(z_i) + W_i, \qquad \text{for } i = 1, \ldots, n, \tag{16}$$

where $Y_i \in \mathbb{R}$ is the observed response variable, $z_i \in \mathcal{Z}$ is a covariate, and W_i is the unobserved error. The errors are assumed to be independent random variables with expectation

$\mathbb{E}W_i = 0$ and variance $\mathrm{Var}(W_i) \le \sigma_0^2 < \infty$, for $i = 1, \ldots, n$. The covariates $z_1, \ldots, z_n$ are fixed, i.e., we consider the case of fixed design.

The function $g_0 : \mathcal{Z} \to \mathbb{R}$ is unknown, but we assume that $g_0 \in \mathcal{G}$, where $\mathcal{G}$ is a given class of regression functions. The unknown regression function can be estimated by the least squares estimator (LSE) $\hat{g}_n$, which is defined (not necessarily uniquely) by

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^n (Y_i - g(z_i))^2. \tag{17}$$

Let

$$Q_n := \frac{1}{n} \sum_{i=1}^n \delta_{z_i}$$

denote the empirical measure of the design points. For $g : \mathcal{Z} \to \mathbb{R}$, we write

$$\|g\|_n^2 := \frac{1}{n}\sum_{i=1}^n g^2(z_i), \quad \|Y - g\|_n^2 := \frac{1}{n}\sum_{i=1}^n (Y_i - g(z_i))^2, \text{ and } \quad \langle W, g\rangle_n := \frac{1}{n}\sum_{i=1}^n W_i g(z_i).$$

Question: When can we say that $\|\hat{g}_n - g_0\|_n \xrightarrow{\mathbb{P}} 0$?

Our starting point is the following inequality:

$$\|\hat{g}_n - g_0\|_n^2 \le 2\langle W, \hat{g}_n - g_0\rangle_n, \tag{18}$$

which follows from simplifying the inequality $\|Y - \hat{g}_n\|_n^2 \le \|Y - g_0\|_n^2$ (Hint: write $\|Y - \hat{g}_n\|_n^2 = \|Y - g_0 + g_0 - \hat{g}_n\|_n^2$ and expand).

We shall need to control the entropy, not of the whole class $\mathcal{G}$ itself, but of subclasses $\mathcal{G}_n(R)$, which are defined as

$$\mathcal{G}_n(R) = \{g \in \mathcal{G} : \|g - g_0\|_n \le R\}.$$

Thus, $\mathcal{G}_n(R)$ is the ball of radius R around g_0, intersected with $\mathcal{G}$.

Theorem 3.20. *Suppose that*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(W_i^2 \mathbf{1}_{\{|W_i| > K\}}\Big) = 0, \tag{19}$$

and

$$\frac{\log N(\delta, \mathcal{G}_n(R), L_1(Q_n))}{n} \to 0, \qquad \textit{for all } \delta > 0, R > 0. \tag{20}$$

Then, $\|\hat{g}_n - g_0\|_n \xrightarrow{p} 0$.

Proof. Let $\eta, \delta > 0$ be given. We will show that $\mathbb{P}(\|\hat{g}_n - g_0\|_n > \delta)$ can be made arbitrarily small, for all n sufficiently large. Note that for any $R > \delta$, we have

$$\mathbb{P}(\|\hat{g}_n - g_0\|_n > \delta) \le \mathbb{P}(\delta < \|\hat{g}_n - g_0\|_n < R) + \mathbb{P}(\|\hat{g}_n - g_0\|_n > R).$$

We will first show that the second term on the right side can be made arbitrarily small by choosing R large. From (18), using Cauchy-Schwarz inequality, it follows that

$$\|\hat{g}_n - g_0\|_n \leq 2\Big(\frac{1}{n}\sum_{i=1}^n W_i^2\Big)^{1/2}.$$

Thus, using Markov's inequality,

$$\mathbb{P}\Big(\|\hat{g}_n - g_0\|_n > R\Big) \leq \mathbb{P}\left(2\Big(\frac{1}{n}\sum_{i=1}^n W_i^2\Big)^{1/2} > R\right) \leq \frac{4}{R^2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}W_i^2 \leq \frac{4\sigma_0^2}{R^2} = \eta,$$

where $R^2 := 4\sigma_0^2/\eta$. Now, using (18) again,

$$\begin{aligned}\mathbb{P}\Big(\delta < \|\hat{g}_n - g_0\|_n < R\Big) \ &\leq \ \mathbb{P}\left(\sup_{g\in\mathcal{G}_n(R)} 2\langle W, g-g_0\rangle_n \geq \delta^2\right)\\ \leq \mathbb{P}\left(\sup_{g\in\mathcal{G}_n(R)} \langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n \geq \frac{\delta^2}{4}\right) &+ \mathbb{P}\left(\sup_{g\in\mathcal{G}_n(R)} \langle W\mathbf{1}_{\{|W|>K\}}, g-g_0\rangle_n \geq \frac{\delta^2}{4}\right). \end{aligned} \tag{21}$$

An application of the Cauchy-Schwarz and Markov's inequality bounds the second term on the right side of the above display:

$$\mathbb{P}\left(\Big(\frac{1}{n}\sum_{i=1}^n W_i^2\mathbf{1}_{\{|W_i|>K\}}\Big)^{1/2} \geq \frac{\delta^2}{4R}\right) \leq \left(\frac{4R}{\delta^2}\right)^2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n W_i^2\mathbf{1}_{\{|W_i|>K\}}\right) \leq \eta,$$

by choosing $K = K(\delta,\eta)$ sufficiently large and using (19). We bound the first term in (21) by using Markov's inequality:

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}_n(R)} \langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n \geq \frac{\delta^2}{4}\right) \leq \frac{4}{\delta^2}\mathbb{E}\left\|\langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n\right\|_{\mathcal{G}_n(R)}.$$

The random variables $W_i\mathbf{1}_{\{|W_i|\leq K\}}$ still have expectation zero if each W_i is symmetric, which we shall assume to avoid digressions. If they are not symmetric, one can use different truncation levels to the left and right to approximately maintain zero expectation. We will now use Hoeffding's inequality (see Lemma 3.9). Thus, for any function $g : \mathcal{Z} \to \mathbb{R}$ and for all $\delta > 0$,

$$\mathbb{P}\Big(|\langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n| \geq \delta\Big) \leq 2\exp\left[-\frac{n\delta^2}{2K^2\|g-g_0\|_n^2}\right]. \tag{22}$$

The proof will now mimic the proof of Theorem 3.5. If $\tilde{\mathcal{G}}$ is an ϵ-net in $L_1(Q_n)$ over $\mathcal{G}_n(R)$, then

$$\mathbb{E}\left\|\langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n\right\|_{\mathcal{G}_n(R)} \leq \mathbb{E}\left\|\langle W\mathbf{1}_{\{|W|\leq K\}}, g-g_0\rangle_n\right\|_{\tilde{\mathcal{G}}} + K\epsilon. \tag{23}$$

The cardinality of $\tilde{\mathcal{G}}$ can be chosen equal to $N(\epsilon, \mathcal{G}_n(R), L_1(Q_n))$. Using the maximal inequality in Lemma 3.13 with $\psi(x) = \exp(x^2)$ (note that (13) holds for every $g \in \tilde{\mathcal{G}}$ by

Lemma 3.21[37] and (22)) shows that the preceding display is does not exceed a multiple of

$$\sqrt{\log N(\epsilon, \mathcal{G}_n(R), L_1(Q_n))}\frac{K}{\sqrt{n}} \sup_{g\in\tilde{\mathcal{G}}} \|g - g_0\|_n + K\epsilon.$$

The norms of $g - g_0 \in \tilde{\mathcal{G}}$ are bounded above by R. By assumption the square root of the entropy divided by $\sqrt{n}$ tends to zero. Thus the above display is less than $(K + 1)\epsilon$ for all large n. Since this argument is valid for every $\epsilon > 0$, it follows that the left side of (23) can be made less than η. □

Exercise (HW1): Assume the setup of (16) where $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} \mid g \text{ in nondecreasing}\}$. Assume that $g_0 \in \mathcal{G}$ is fixed, and $\sup_{z\in\mathbb{R}} |g_0(z)| \le K$ for some (unknown) constant K. Using the fact that

$$\log N_{[\,]}(\epsilon, \mathcal{G}_{[0,1]}, L_1(Q_n)) \le A\epsilon^{-1}, \qquad \text{for all } \epsilon > 0,$$

where $\mathcal{G}_{[0,1]} = \{g : \mathbb{R} \to [0, 1] \mid g \text{ in nondecreasing}\}$ and $A > 0$ is a universal constant, show that $\|\hat{g}_n - g_0\|_n \xrightarrow{p} 0$ (here $\hat{g}_n$ is the LSE defined in (17)).

3.6 Bounded differences inequality — a simple concentration inequality

We are interested in bounding the random fluctuations of (complicated) functions of many independent random variables (e.g., the Hoeffding's inequality in Lemma 3.9 accomplishes this for the sum of independent bounded random variables). Let $X_1, \ldots, X_n$ be *independent* random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$, and let

$$Z = f(X_1, \ldots, X_n)$$

be the random variable of interest (e.g., $Z = \sum_{i=1}^n X_i$ when $\mathcal{X} = \mathbb{R}$). We seek upper bounds for

$$\mathbb{P}(Z > \mathbb{E}Z + t) \qquad \text{and} \qquad \mathbb{P}(Z < \mathbb{E}Z - t) \qquad \text{for } t > 0.$$

In this subsection we study a concentration inequality similar in spirit to the Hoeffding's inequality, but that holds for a general random variable $Z = f(X_1, \ldots, X_n)$, under suitable conditions.

---

[37] The following result shows that sub-Gaussian tail bounds for a random variable W imply the finiteness of $\mathbb{E}e^{DW^2}$ for some $D > 0$.

Lemma 3.21. *Let W be a random variable with $\mathbb{P}(|W| > x) \le Ke^{-Cx^2}$ for every x, for constants K and C. Then, $\mathbb{E}e^{DW^2} \le KD/(C - D) + 1$, for any $D < C$.*

Proof. By Fubini's theorem,

$$\mathbb{E}(e^{DW^2} - 1) = \mathbb{E}\int_0^{|W|^2} De^{Ds}ds = \int_0^\infty \mathbb{P}(|W| > \sqrt{s})De^{Ds}ds \le KD\int_0^\infty e^{-Cs}e^{Ds}ds = \frac{KD}{C - D}.$$

□

Let $X_1, \ldots, X_n$ be *independent* random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and

$$Z = f(X_1, \ldots, X_n)$$

be the random variable of interest. Note that if we define

$$Y_k := \mathbb{E}[Z|X_1, \ldots, X_k], \qquad \text{for } k = 1, \ldots, n,$$

then $\{Y_k\}_{k=0}^n$ is a martingale[38] adapted to a filtration generated by $\{X_k\}_{k=1}^n$.

Denote by $\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot|X_1, \ldots, X_i]$. Thus, $\mathbb{E}_0(Z) = \mathbb{E}(Z)$ and $\mathbb{E}_k(Z) = Y_k$, for $k = 1, \ldots, n$. Writing

$$\Delta_i := \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z],$$

we have

$$Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i.$$

We want to get exponential concentration inequalities for Z. As before, we start with bounding the moment generating function $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$. We start with the Azuma-Hoeffding inequality for sums of bounded martingale differences.

Lemma 3.22 (Azuma-Hoeffding inequality)**.** *Suppose that the martingale differences are bounded, i.e., $|\Delta_i| \le c_i$, for all $i = 1, \ldots, n$. Then,*

$$\mathbb{E}e^{\lambda(Z-\mathbb{E}(Z))} \le e^{\lambda^2 \sum_{i=1}^n c_i^2/2}.$$

Proof. Observe that,

$$\begin{aligned}
\mathbb{E}e^{\lambda(Z-\mathbb{E}(Z))} &= \mathbb{E}e^{\lambda \sum_{i=1}^n \Delta_i} = \mathbb{E}\left[\mathbb{E}_{n-1}(e^{\lambda(\sum_{i=1}^{n-1}\Delta_i)+\lambda\Delta_n})\right] \\
&= \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1}\Delta_i)}\right]\mathbb{E}_{n-1}[e^{\lambda\Delta_n}] \\
&\le \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1}\Delta_i)}\right]e^{\lambda^2 c_n^2/2} \qquad \text{(by Lemma 3.8)} \\
&\cdots \\
&\le e^{\lambda^2(\sum_{i=1}^n c_i^2)/2}.
\end{aligned}$$

□

Definition 3.23 (Functions with bounded differences)**.** *We say that a function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded difference property if for some nonnegative constants $c_1, \ldots, c_n$,*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i, \quad 1 \le i \le n. \tag{24}$$

[38] Given a sequence $\{Y_k\}_{k=1}^\infty$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^\infty$ (e.g., $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$), the pair $\{Y_k, \mathcal{F}_k\}_{k=1}^\infty$ is a martingale if, for all $k \ge 1$,

$$\mathbb{E}[|Y_k|] < \infty, \quad \text{and} \quad \mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k.$$

In other words, if we change the i-th variable of f while keeping all the others fixed, the value of the function cannot change by more than c_i.

The following theorem provides exponential tail bounds for the random variable $Z - \mathbb{E}(Z)$. It follows easily from the above lemma and is left as an exercise.

Theorem 3.24 (Bounded differences inequality or McDiarmid's inequality)**.** *Suppose that $Z = f(X_1, \ldots, X_n)$ and f is such that (24) holds (i.e., f has the bounded differences property), then*

$$\mathbb{P}(|Z - \mathbb{E}(Z)| > t) \leq 2e^{-2t^2/\sum_{i=1}^n c_i^2}. \tag{25}$$

Proof. We will show that if f satisfies the bounded difference property with the constants c_i's, then $|\Delta_i| \leq c_i$, which together with Lemma 3.22 will yield the desired result. Recall that

$$\Delta_i \equiv \Delta_i(X_1, \ldots, X_i) := \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i] - \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}].$$

Fix $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$, for any $x_1, \ldots, x_{i-1} \in \mathcal{X}$. Then Δ_i can be viewed solely as a function of X_i. We will study the range of Δ_i as $X_i = x$ varies over $\mathcal{X}$. Then, observe that, as $X_1, \ldots, X_n$ are independent,

$$\begin{aligned}
& |\Delta_i(x_1, \ldots, x_{i-1}, x)| \\
= & \left|\mathbb{E}[f(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n)] - \mathbb{E}[f(x_1, \ldots, x_{i-1}, X_i, X_{i+1}, \ldots, X_n)]\right| \\
= & \left|\mathbb{E}\big[f(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - f(x_1, \ldots, x_{i-1}, X_i, X_{i+1}, \ldots, X_n)\big]\right| \\
\leq & \ \mathbb{E}\left[\left|f(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - f(x_1, \ldots, x_{i-1}, X_i, X_{i+1}, \ldots, X_n)\right|\right] \\
\leq & \ \mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} \left|f(x_1, \ldots, x_{i-1}, x, X_{i+1}, \ldots, X_n) - f(x_1, \ldots, x_{i-1}, x', X_{i+1}, \ldots, X_n)\right|\right] \\
\leq & \ c_i
\end{aligned}$$

where the first inequality follows using Jensen's inequality, and the last inequality follows from the fact that f satisfies the bounded differences property (24). $\square$

Theorem 3.24 can be seen as a quantification of the following qualitative statement of Talagrand (see [Talagrand, 1996b, Page 2]): *A random variable that depends on the influence of many independent variables (but not too much on any of them) concentrates.* The numbers c_i control the effect of the i-th variable on the function f.

The bounded differences inequality is quite useful and distribution-free and often close to optimal. However, it does not incorporate the variance information.

Example 3.25 (Kernel density estimation)**.** *Let $X_1, \ldots, X_n$ are i.i.d. from a distribution P on $\mathbb{R}$ (the argument can be easily generalized to $\mathbb{R}^d$) with density ϕ. We want to estimate*

ϕ nonparametrically using the kernel density estimator (KDE) $\hat{\phi}_n : \mathbb{R} \to [0, \infty)$ defined as

$$\hat{\phi}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right), \qquad for x \in \mathbb{R},$$

where $h_n > 0$ is the smoothing bandwidth and K is a nonnegative kernel (i.e., $K \geq 0$ and $\int K(x)dx = 1$). The L_1-error of the estimator $\hat{\phi}_n$ is

$$Z \equiv f(X_1, \ldots, X_n) := \int |\hat{\phi}_n(x) - \phi(x)| dx.$$

The random variable Z not only provides a measure of the difference between $\hat{\phi}_n$ and ϕ but, as $Z = 2 \sup_A |P_n(A) - P(A)|$ (Exercise (HW1): Show this) where the supremum is over all Borel sets in $\mathbb{R}$ and P_n denotes the distribution corresponding to the KDE $\hat{\phi}_n$), Z also captures the difference between P_n and P in the total variation distance.

We can use Theorem 3.24 to get exponential tail bounds for Z. We will show that (25) holds with $c_i = 2/n$, for all $i = 1, \ldots, n$. It is easy to see that for $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$\begin{aligned} & |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \\ \leq \;& \frac{1}{nh_n} \int \left| K\left(\frac{x - x_i}{h_n}\right) - K\left(\frac{x - x_i'}{h_n}\right) \right| dx \leq \frac{2}{n}. \end{aligned} \tag{26}$$

Thus, using Theorem 3.24 we have

$$\mathbb{P}(|Z - \mathbb{E}(Z)| > t) \leq 2e^{-nt^2/2} \qquad \Rightarrow \qquad P(\sqrt{n}|Z - \mathbb{E}(Z)| > t) \leq 2e^{-t^2/2},$$

which shows that Z concentrates around its expectation $\mathbb{E}[Z]$ at the rate $n^{-1/2}$. The remarkable thing is that this concentration property holds regardless of the choice of bandwidth h_n. Of course, in this case, it is difficult to actually compute what that expectation is.

3.7 Supremum of the empirical process for a bounded class of functions

Let us conclude this section with an important result on what the bounded differences concentration inequality implies for the supremum of the empirical process

$$Z := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \right|,$$

where $X_1, \ldots, X_n$ are i.i.d. random objects taking values in $\mathcal{X}$ and $\mathcal{F}$ is a collection of real-valued functions on $\mathcal{X}$, when it is assumed that all functions in $\mathcal{F}$ are bounded by a positive constant B, i.e.,

$$\sup_{x \in \mathcal{X}} |f(x)| \leq B \qquad \text{for all } f \in \mathcal{F}.$$

We shall argue that Z concentrates around its expectation. Let

$$g(x_1, \ldots, x_n) := \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}[f(X_1)] \right|.$$

We shall show below that g satisfies the bounded differences property (24) with $c_i := 2B/n$ for $i = 1, \ldots, n$. To see this, note that

$$\begin{aligned} g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) &= \left| \frac{1}{n} \sum_{j \neq i} f(x_i) + \frac{f(x_i')}{n} - \mathbb{E}[f(X_1)] \right| \\ &= \left| \frac{1}{n} \sum_{j=1}^n f(x_j) - \mathbb{E}[f(X_1)] + \frac{f(x_i')}{n} - \frac{f(x_i)}{n} \right| \\ &\leq \left| \frac{1}{n} \sum_{j=1}^n f(x_j) - \mathbb{E}[f(X_1)] \right| + \frac{2B}{n} \quad \leq \quad g(x_1, \ldots, x_n) + \frac{2B}{n}, \end{aligned}$$

where we have used the fact that for every $f \in \mathcal{F}$, $|f(x_i)| \leq B$ and $f(x_i')| \leq B$. Interchanging the roles of x_i and x_i', we can deduce that (24) holds with $c_i := 2B/n$ for $i = 1, \ldots, n$. Then, Theorem 3.24 yields

$$\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq 2 \exp\left(-\frac{nt^2}{2B^2}\right), \qquad \text{for every } t \geq 0.$$

Setting $\delta := \exp\left(-\frac{nt^2}{2B^2}\right)$, we can deduce that

$$|Z - \mathbb{E}[Z]| \leq B\sqrt{\frac{2}{n} \log \frac{1}{\delta}},$$

holds with probability at least $1 - 2\delta$ for every $\delta > 0$. This inequality implies that $\mathbb{E}[Z]$ is usually the dominating term for understanding the behavior of Z.

We may apply this to study the classical Glivenko-Cantelli problem. The following theorem illustrates this.

Theorem 3.26. *Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables on $\mathbb{R}$ with distribution P and c.d.f. F. Let $\mathbb{F}_n$ be the empirical d.f. of the data (see (1)). Then,*

$$\mathbb{P}\left[\|\mathbb{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + t\right] \leq e^{-nt^2/2}, \qquad \text{forall } \ t > 0. \tag{27}$$

Hence, $\|\mathbb{F}_n - F\|_\infty \xrightarrow{a.s.} 0$.

Proof. The function class under consideration is $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$. Then, $Z := \|\mathbb{P}_n - P\|_{\mathcal{F}} = \|\mathbb{F}_n - F\|_\infty$. From the discussion in this subsection, we have to bound upper bound $\mathbb{E}[Z]$. This can be done via symmetrization, i.e., $\mathbb{E}[Z] \leq 2\mathbb{E}_X[\mathbb{E}_\varepsilon[\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)|]]$, where $\varepsilon_1, \ldots, e_n$ are i.i.d. Rademachers independent of the X_i's. For a fixed $(x_1, \ldots, x_n) \in \mathbb{R}^n$, define

$$\Delta_n(\mathcal{F}; x_1, \ldots, x_n) := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}.$$

Observe that although $\mathcal{F}$ has uncountable many functions, for every $(x_1, \ldots, x_n) \in \mathbb{R}^n$, $\Delta_n(\mathcal{F}; x_1, \ldots, x_n)$ can take at most $n + 1$ distinct values[39]. Thus, $\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|$

[39] If we order x_1^n as $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, then they split the real line into at most $n + 1$ intervals (including the two end-intervals $(-\infty, x_{(1)})$ and $[x_{(n)}, \infty)$. Thus, for a given $t \in \mathbb{R}$, the indicator $\mathbf{1}_{(-\infty,t]}(x_{(i)})$ takes the value one for all $x_{(i)} \leq t$, and the value zero for all other samples.

is at most the supremum of $n+1$ such variables, and we can apply Lemma 3.16 to show that[40]

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\right] \le 4\sqrt{\frac{\log(n+1)}{n}}.$$

This shows (27).

Exercise (HW2): Show that this implies that $\|\mathbb{F}_n - F\|_\infty \overset{a.s.}{\to} 0$. $\square$

Although the exponential tail bound (27) is adequate for many purposes, it is far from the tightest possible. Using alternative methods (using Dudley's entropy bound in Section 4), we provide a sharper result that removes the $\log(n+1)$ factor.

[40] Note that for $x_1,\ldots,x_n$ distinct

$$\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big| \overset{d}{=} \max_{j=1,\ldots,n}\Big|\frac{1}{n}\sum_{i=1}^{j}\varepsilon_i\Big|$$

and direct calculations would actually show that the $\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\right] = O(n^{-1/2})$.

4 Chaining and uniform entropy

In this section we will introduce Dudley's metric entropy bound (and the idea of chaining). We will use this result to prove a maximal inequality (with uniform entropy) that will be useful in deriving rates of convergence of statistical estimators (see Section 5). Further, as we will see later, these derived maximal inequalities also play a crucial role in proving functional extensions of the Donsker's theorem (see Section 11). In fact, these maximal inequalities are at the heart of the theory of empirical processes.

The proof of the main result involves in this section an idea called chaining. Before we start with chaining, let us recall our first maximal inequality (13) (see Lemma 3.13). Note that the bound (13) can be tight in some situations. For example, this is the case when $Y_1, \ldots, Y_N$ are i.i.d. $N(0, \sigma^2)$. Because of this example, Lemma 3.13 cannot be improved without imposing additional conditions on the $Y_1, \ldots, Y_N$. It is also easy to construct examples where (13) is quite weak. For example, if $Y_i = Y_0 + \sigma_N Z_i$, for $i = 1, \ldots, N$, for some $Y_0 \sim N(0, \sigma^2)$ and $Z_i \stackrel{iid}{\sim} N(0,1)$ and $\sigma_N \sqrt{\log N} = o(1)$, then it is clear that $\max_{i=1,\ldots,N} |Y_i| \approx Y_0$ so that (13) will be loose by a factor of $\sqrt{\log N}$. In order to improve on (13), we need to make assumptions on how *close* to each other the Y_i's are. Dudley's entropy bound makes such an assumption explicit and provides improved upper bounds for $\mathbb{E}[\max_{i=1,\ldots,N} |Y_i|]$.

4.1 Dudley's bound for the supremum of a sub-Gaussian process

For generality, we will assume that we may have an infinite (possibly uncountable) collection of random variables and we are interested in the expected supremum of the collection.

Suppose that (T, d) is a metric space and X_t is a stochastic process indexed by T. We will state two versions of Dudley's metric entropy bound and will prove one of these results (the other has a similar proof). Let us first assume that

$$\mathbb{E} X_t = 0, \qquad \text{for all } \ t \in T.$$

We want to find upper bounds for

$$\mathbb{E} \sup_{t \in T} X_t$$

that ONLY depends on structure of the metric space (T, d). We shall first state Dudley's bound when the index set T is finite and subsequently improve it to the case when T is infinite.

Theorem 4.1 (Dudley's entropy bound for finite T)**.** *Suppose that $\{X_t : t \in T\}$ is a mean zero stochastic process such that for every $s, t \in T$ and $u \geq 0$,*[41]

$$\mathbb{P}\left\{|X_t - X_s| \geq u\right\} \leq 2 \exp\left(-\frac{u^2}{2d^2(s,t)}\right). \tag{28}$$

[41]Note that if $X_t, t \in T$, have mean zero and are jointly Gaussian, then $X_t - X_s$ is a mean zero normal random variable for every $s, t \in T$ so that (28) holds with $d(s,t) := \sqrt{\mathbb{E}(X_s - X_t)^2}$.

Also, assume that (T, d) is a finite metric space. Then, we have

$$\mathbb{E}\Big[\sup_{t\in T} X_t\Big] \leq C \int_0^\infty \sqrt{\log N(\epsilon, T, d)}\, d\epsilon \tag{29}$$

where $C > 0$ is a constant.

Next we give a slightly different formulation of Dudley's metric entropy bound for finite T. However, before proceeding further, we shall give a result similar to Lemma 3.13 but instead bound the maximum of the absolute values of sub-Gaussian random variables.

Proposition 4.2. *Let T be a finite set and let $\{X_t, t \in T\}$ be a stochastic process. Suppose that for every $t \in T$ and $u \geq 0$, the inequality*

$$\mathbb{P}(|X_t| \geq u) \leq 2 \exp\left(-\frac{u^2}{2\sigma^2}\right) \tag{30}$$

holds. Here σ is a fixed positive real number. Then, for a universal positive constant C, we have

$$\mathbb{E} \max_{t\in T} |X_t| \leq C\sigma\sqrt{\log(2|T|)}. \tag{31}$$

Proof of Proposition 4.2. Because

$$\mathbb{E} \max_{t\in T} |X_t| = \int_0^\infty \mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) du,$$

we can control $\mathbb{E}\max_{t\in T} |X_t|$ by bounding the tail probability $\mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) du$ for every $u \geq 0$. For this, write

$$\mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) = \mathbb{P}\left(\cup_{t\in T}\{|X_t| \geq u\}\right) \leq \sum_{t\in T} \mathbb{P}\left(|X_t| \geq u\right) \leq 2|T| \exp\left(-\frac{u^2}{2\sigma^2}\right).$$

This bound is good for large u but not so good for small u (it is quite bad for $u = 0$ for example). It is therefore good to use it only for $u \geq u_0$ for some u_0 to be specified later. This gives[42]

$$\begin{aligned}
\mathbb{E} \max_{t\in T} |X_t| &= \int_0^\infty \mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) du \\
&= \int_0^{u_0} \mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) du + \int_{u_0}^\infty \mathbb{P}\left(\max_{t\in T} |X_t| \geq u\right) du \\
&\leq u_0 + \int_{u_0}^\infty 2|T| \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\
&\leq u_0 + \int_{u_0}^\infty 2|T| \frac{u}{u_0} \exp\left(-\frac{u^2}{2\sigma^2}\right) du = u_0 + \frac{2|T|}{u_0}\sigma^2 \exp\left(-\frac{u_0^2}{2\sigma^2}\right).
\end{aligned}$$

[42]**Result**: Suppose that $Z \geq 0$. Then, $\mathbb{E}(Z) = \int_0^\infty \mathbb{P}(Z \geq t)\, dt$.

One can try to minimize the above term over u_0. A simpler strategy is to realize that the large term here is $2|T|$ so one can choose u_0 to kill this term by setting

$$\exp\left(\frac{u_0^2}{2\sigma^2}\right) = 2|T| \quad \text{or} \quad u_0 = \sqrt{2}\sigma\sqrt{\log(2|T|)}.$$

This gives

$$\mathbb{E}\max_{t\in T}|X_t| \le \sqrt{2}\sigma\sqrt{\log(2|T|)} + \frac{\sigma^2}{\sqrt{2\sigma^2\log(2|T|)}} \le C\sigma\sqrt{\log(2|T|)}$$

which proves the result. □

Theorem 4.3. *Suppose (T,d) is a finite metric space and $\{X_t, t\in T\}$ is a stochastic process such that (28) hold. Then, for a universal positive constant C, the following inequality holds for every $t_0\in T$:*

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le C\int_0^\infty \sqrt{\log D(\epsilon,T,d)}\,d\epsilon \lesssim \int_0^\infty \sqrt{\log N(\epsilon,T,d)}\,d\epsilon. \tag{32}$$

Here $D(\epsilon,T,d)$ denotes the ϵ-packing number of the space (T,d).

The following remarks mention some alternative forms of writing the inequality (32) and also describe some implications.

Remark 4.1. *Let $\tilde{D}$ denote the diameter of the metric space T (i.e., $\tilde{D} = \max_{s,t\in T} d(s,t)$). Then the packing number $D(\epsilon,T,d)$ clearly equals 1 for $\epsilon \ge \tilde{D}$ (it is impossible to have two points in T whose distance is strictly larger than ϵ when $\epsilon > D$). Therefore,*

$$\int_0^\infty \sqrt{\log D(\epsilon,T,d)}d\epsilon = \int_0^{\tilde{D}} \sqrt{\log D(\epsilon,T,d)}d\epsilon.$$

Moreover,

$$\begin{aligned}\int_0^{\tilde{D}} \sqrt{\log D(\epsilon,T,d)}d\epsilon &= \int_0^{\tilde{D}/2} \sqrt{\log D(\epsilon,T,d)}d\epsilon + \int_{\tilde{D}/2}^{\tilde{D}} \sqrt{\log D(\epsilon,T,d)}d\epsilon \\ &\le \int_0^{\tilde{D}/2} \sqrt{\log D(\epsilon,T,d)}d\epsilon + \int_0^{\tilde{D}/2} \sqrt{\log D(\epsilon+(\tilde{D}/2),T,d)}d\epsilon \\ &\le 2\int_0^{\tilde{D}/2} \sqrt{\log D(\epsilon,T,d)}d\epsilon\end{aligned}$$

because $D(\epsilon+(\tilde{D}/2),T,d) \le D(\epsilon,T,d)$ for every ϵ. We can thus state Dudley's bound as

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le C\int_0^{\tilde{D}/2} \sqrt{\log D(\epsilon,T,d)}d\epsilon$$

where the C above equals twice the constant C in (32). Similarly, again by splitting the above integral in two parts (over 0 to $\tilde{D}/4$ and over $\tilde{D}/4$ to $\tilde{D}/2$), we can also state Dudley's bound as

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le C\int_0^{\tilde{D}/4} \sqrt{\log D(\epsilon,T,d)}d\epsilon.$$

The constant C above now is 4 times the constant in (32).

Remark 4.2. *The left hand side in* (32) *is bounded from below (by triangle inequality) by* $\mathbb{E}\max_{t\in T}|X_t| - \mathbb{E}|X_{t_0}|$*. Thus,* (32) *implies that*

$$\mathbb{E}\max_{t\in T}|X_t| \le \mathbb{E}|X_{t_0}| + C\int_0^{\tilde{D}/4}\sqrt{\log D(\epsilon, T, d)}\,d\epsilon \qquad \textit{for every } t_0 \in T.$$

We shall now give the proof of Theorem 4.3. The proof will be based on an idea called *chaining*. Specifically, we shall split $\max_{t\in T}(X_t - X_{t_0})$ in chains and use the bound given by Proposition 4.2 within the links of each chain.

Proof of Theorem 4.3. Recall that $\tilde{D}$ is the diameter of T. For $n \ge 1$, let T_n be a maximal $\tilde{D}2^{-n}$-separated subset of T i.e., $\min_{s,t\in T_n: s\ne t} d(s,t) > \tilde{D}2^{-n}$ and T_n has maximal cardinality subject to the separation restriction. The cardinality of T_n is given by the packing number $D(\tilde{D}2^{-n}, T, d)$. Because of the maximality,

$$\max_{t\in T}\min_{s\in T_n} d(s,t) \le \tilde{D}2^{-n}. \tag{33}$$

Because T is finite and $d(s,t) > 0$ for all $s \ne t$, the set T_n will equal T when n is large. Let

$$N := \min\{n \ge 1 : T_n = T\}.$$

For each $n \ge 1$, let $\pi_n : T \to T_n$ denote the function which maps each point $t \in T$ to the point in T_n that is closest to T (if there are multiple closest points to T in T_n, then choose one arbitrarily). In other words, $\pi_n(t)$ is chosen so that

$$d(t, \pi_n(t)) = \min_{s\in T_n} d(t,s).$$

As a result, from (33), we have

$$d(t, \pi_n(t)) \le \tilde{D}2^{-n} \qquad \text{for all } t \in T \text{ and } n \ge 1. \tag{34}$$

Note that $\pi_N(t) = t$. Finally let $T_0 := \{t_0\}$ and $\pi_0(t) = t_0$ for all $t \in T$.

We now note that

$$X_t - X_{t_0} = \sum_{n=1}^{N}\left(X_{\pi_n(t)} - X_{\pi_{n-1}(t)}\right) \qquad \text{for every } t \in T. \tag{35}$$

The sequence

$$t_0 \to \pi_1(t) \to \pi_2(t) \to \cdots \to \pi_{N-1}(t) \to \pi_N(t) = t$$

can be viewed as a chain from t_0 to t. This is what gives the argument the name *chaining*.

By (35), we obtain

$$\max_{t\in T}|X_t - X_{t_0}| \le \max_{t\in T}\sum_{n=1}^{N}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \le \sum_{n=1}^{N}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|$$

so that

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le \sum_{n=1}^{N}\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|. \tag{36}$$

Now to bound $\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|$ for each $1 \le n \le N$, we shall use the elementary bound given by Proposition 4.2. For this, note first that by (28), we have

$$\mathbb{P}\left\{|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \ge u\right\} \le 2\exp\left(\frac{-u^2}{2d^2(\pi_n(t), \pi_{n-1}(t))}\right).$$

Now

$$d(\pi_n(t), \pi_{n-1}(t)) \le d(\pi_n(t), t) + d(\pi_{n-1}(t), t) \le \tilde{D}2^{-n} + \tilde{D}2^{-(n-1)} = 3\tilde{D}2^{-n}.$$

Thus Proposition 4.2 can be applied with $\sigma := 3\tilde{D}2^{-n}$ so that we obtain (**note that the value of C might change from occurrence to occurrence**)

$$\begin{aligned}\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| &\le C\frac{3\tilde{D}}{2^n}\sqrt{\log\left(2|T_n||T_{n-1}|\right)}\\ &\le C\tilde{D}2^{-n}\sqrt{\log\left(2|T_n|^2\right)} \le C\tilde{D}2^{-n}\sqrt{\log\left(2D(\tilde{D}2^{-n}, T, d)\right)}\end{aligned}$$

Plugging the above bound into (36), we deduce

$$\begin{aligned}\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| &\le C\sum_{n=1}^{N}\frac{\tilde{D}}{2^n}\sqrt{\log\left(2D(\tilde{D}2^{-n}, T, d)\right)}\\ &\le 2C\sum_{n=1}^{N}\int_{\tilde{D}/2^{n+1}}^{\tilde{D}/2^n}\sqrt{\log(2D(\epsilon, T, d))}d\epsilon\\ &= C\int_{\tilde{D}/2^{N+1}}^{\tilde{D}/2}\sqrt{\log(2D(\epsilon, T, d))}d\epsilon\\ &\le C\int_0^{\tilde{D}/2}\sqrt{\log(2D(\epsilon, T, d))}d\epsilon\\ &= C\int_0^{\tilde{D}/4}\sqrt{\log(2D(\epsilon, T, d))}d\epsilon + C\int_0^{\tilde{D}/4}\sqrt{\log(2D(\epsilon + (\tilde{D}/4), T, d))}d\epsilon\\ &\le 2C\int_0^{\tilde{D}/4}\sqrt{\log(2D(\epsilon, T, d))}d\epsilon.\end{aligned}$$

Note now that for $\epsilon \le \tilde{D}/4$, the packing number $D(\epsilon, T, d) \ge 2$ so that

$$\log(2D(\epsilon, T, d)) \le \log 2 + \log D(\epsilon, T, d) \le 2\log D(\epsilon, T, d).$$

We have thus proved that

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le 2\sqrt{2}C\int_0^{\tilde{D}/4}\sqrt{\log D(\epsilon, T, d)}d\epsilon$$

which proves (32). $\square$

4.1.1 Dudley's bound when the metric space is separable

We shall next prove Dudley's bound for the case of infinite T. This requires a technical assumption called *separability* which will always be satisfied in our applications.

Definition 4.4 (Separable stochastic process)**.** *Let (T, d) be a metric space. The stochastic process $\{X_t, t \in T\}$ indexed by T is said to be separable if there exists a null set N and a countable subset $\tilde{T}$ of T such that for all $\omega \notin N$ and $t \in T$, there exists a sequence $\{t_n\}$ in $\tilde{T}$ with $\lim_{n\to\infty} d(t_n, t) = 0$ and $\lim_{n\to\infty} X_{t_n}(\omega) = X_t(\omega)$.*

Note that the definition of separability requires that $\tilde{T}$ is a dense subset of T which means that the metric space (T, d) is separable (a metric space is said to be separable if it has a countable dense subset).

The following fact is easy to check: If (T, d) is a separable metric space and if $X_t, t \in T$, has continuous sample paths (almost surely), then $X_t, t \in T$ is separable. The statement that $X_t, t \in T$, has continuous sample paths (almost surely) means that there exists a null set N such that for all $\omega \notin N$, the function $t \mapsto X_t(\omega)$ is continuous on T.

The following fact is also easy to check: If $\{X_t, t \in T\}$ is a separable stochastic process, then

$$\sup_{t\in T} |X_t - X_{t_0}| = \sup_{t\in \tilde{T}} |X_t - X_{t_0}| \qquad \text{almost surely} \tag{37}$$

for every $t_0 \in T$. Here $\tilde{T}$ is a countable subset of T which appears in the definition of separability of $X_t, t \in T$.

In particular, the statement (37) implies that $\sup_{t\in T} |X_t - X_{t_0}|$ is measurable (note that uncountable suprema are in general not guaranteed to be measurable; but this is not an issue for separable processes).

We shall now state Dudley's theorem for separable processes. This theorem does not impose any cardinality restrictions on T (it holds for both finite and infinite T).

Theorem 4.5. *Let (T, d) be a separable metric space and let $\{X_t, t \in T\}$ be a separable stochastic process. Suppose that for every $s, t \in T$ and $u \geq 0$, we have*

$$\mathbb{P}\{|X_s - X_t| \geq u\} \leq 2\exp\left(-\frac{u^2}{2d^2(s,t)}\right).$$

Then for every $t_0 \in T$, we have

$$\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \leq C\int_0^{\tilde{D}/4} \sqrt{\log D(\epsilon, T, d)}\, d\epsilon \tag{38}$$

where $\tilde{D}$ is the diameter of the metric space (T, d).

Proof of Theorem 4.5. Let $\tilde{T}$ be a countable subset of T such that (37) holds. We may assume that $\tilde{T}$ contains t_0 (otherwise simply add t_0 to $\tilde{T}$). For each $k \geq 1$, let $\tilde{T}_k$ be the finite set obtained by taking the first k elements of $\tilde{T}$ (in an arbitrary enumeration of the entries of $\tilde{T}$). We can ensure that $\tilde{T}_k$ contains t_0 for every $k \geq 1$.

Applying the finite index set version of Dudley's theorem (Theorem 4.3) to $\{X_t, t \in \tilde{T}_k\}$, we obtain

$$\mathbb{E} \max_{t \in \tilde{T}_k} |X_t - X_{t_0}| \leq C \int_0^{diam(\tilde{T}_k)/4} \sqrt{\log D(\epsilon, \tilde{T}_k, d)} d\epsilon \leq C \int_0^{\tilde{D}/4} \sqrt{\log D(\epsilon, T, d)} d\epsilon.$$

Note that the right hand side does not depend on k. Letting $k \to \infty$ on the left hand side, we use the Monotone Convergence Theorem to obtain

$$\mathbb{E}\left[\sup_{t \in \tilde{T}} |X_t - X_{t_0}|\right] \leq C \int_0^{\tilde{D}/4} \sqrt{\log D(\epsilon, T, d)} d\epsilon.$$

The proof is now completed by (37). $\square$

Remark 4.3. *One may ask if there is a lower bound for $\mathbb{E}\sup_{t \in T} X_t$ in terms of covering/packing numbers. A classical result in this direction is Sudakov's lower bound which states: For a zero-mean Gaussian process X_t defined on T, define the variance pseudometric $d^2(s,t) := \mathrm{Var}(X_s - X_t)$. Then,*

$$\mathbb{E}\Big[\sup_{t \in T} X_t\Big] \geq \sup_{\epsilon > 0} \frac{\epsilon}{2} \sqrt{\log D(\epsilon, T, d)},$$

where $D(\epsilon, T, d)$ is the ϵ-packing number of (T, d).

It is natural to ask how Dudley's bound can be useful for the theory of empirical process. Indeed, Theorem 4.5 is enormously helpful in upper bounding the supremum of the empirical process as indicated by the maximum inequality in the next subsection.

4.2 Maximal inequality with uniform entropy

Recall our setup: We have data $X_1, \ldots, X_n$ i.i.d. P on $\mathcal{X}$ and a class of real valued functions $\mathcal{F}$ defined on $\mathcal{X}$. For any function $f : \mathcal{X} \to \mathbb{R}$,

$$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^n f^2(X_i)$$

denotes the $L_2(\mathbb{P}_n)$-seminorm. Further, recall that the empirical process under consideration is the stochastic process indexed by $\mathcal{F}$ and defined as $\mathbb{G}_n(f) = \sqrt{n}(\mathbb{P}_n - P)(f)$, for $f \in \mathcal{F}$.

Definition 4.6 (Uniform entropy bound)**.** *A class $\mathcal{F}$ of measurable functions with measurable envelope F satisfies the uniform entropy bound if and only if $J(1, \mathcal{F}, F) < \infty$ where*

$$J(\delta, \mathcal{F}, F) := \int_0^\delta \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F} \cup \{0\}, L_2(Q))}\, d\epsilon, \qquad \delta > 0. \tag{39}$$

Here the supremum if taken over all finitely discrete probability measures Q on $\mathcal{X}$ with $\|F\|_{Q,2}^2 := \int F^2 dQ > 0$ and we have added the function $f \equiv 0$ to $\mathcal{F}$. Finiteness of the previous integral will be referred to as the uniform entropy condition.

The uniform entropy integral may seem a formidable object, but we shall later see how to bound it for concrete classes $\mathcal{F}$. Of course, the class $\mathcal{F}$ must be totally bounded in $L_2(Q)$ to make the integrand in the integral bounded, and then still the integral might diverge (evaluate to $+\infty$). The integrand is a nonincreasing function of ϵ, and finiteness of the integral is therefore determined by its behaviour near $\epsilon = 0$. Because $\int_0^1 (1/\epsilon)^r d\epsilon$ is finite if $r < 1$ and infinite if $r \geq 1$, convergence of the integral roughly means that, for $\epsilon \downarrow 0$,

$$\sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) << \left(\frac{1}{\epsilon}\right)^2,$$

where $<<$ means smaller in order, or smaller up to an appropriate logarithmic term.

Theorem 4.7. *If $\mathcal{F}$ is a class of measurable functions with measurable envelope function F, then*

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim \mathbb{E}\Big[J(\theta_n, \mathcal{F}, F)\|F\|_n\Big] \lesssim J(1, \mathcal{F}, F)\|F\|_{P,2}, \tag{40}$$

where $\theta_n := \sup_{f \in \mathcal{F}} \|f\|_n / \|F\|_n$.

Proof. By symmetrization (see Theorem 3.17) it suffices to bound $\mathbb{E}\|\mathbb{G}_n^o\|_{\mathcal{F}}$; recall that $\mathbb{G}_n^o(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i)$ where ε_i's are i.i.d Rademacher. Given $X_1, \ldots, X_n$, the process $\mathbb{G}_n^o$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$-seminorm $\|\cdot\|_n$ (by Lemma 3.12), i.e.,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n \varepsilon_i \frac{f(X_i)}{\sqrt{n}} - \sum_{i=1}^n \varepsilon_i \frac{g(X_i)}{\sqrt{n}}\Big| \geq u \Big| X_1, \ldots, X_n\right) \leq 2e^{-u^2/(2\|f-g\|_n^2)}, \ \forall\, f, g \in \mathcal{F}, \ \forall\, u \geq 0.$$

The value $\sigma_{n,2}^2 := \sup_{f \in \mathcal{F}} \mathbb{P}_n f^2 = \sup_{f \in \mathcal{F}} \|f\|_n^2$ is an upper bound for the squared radius of $\mathcal{F} \cup \{0\}$ with respect to this norm. We add the function $f \equiv 0$ to $\mathcal{F}$, so that the symmetrized process is zero at some parameter. The maximal inequality (38) (with $X_{t_0} = 0$) gives

$$\mathbb{E}_\varepsilon \|\mathbb{G}_n^o\|_{\mathcal{F}} \lesssim \int_0^{\sigma_{n,2}} \sqrt{\log N(\epsilon, \mathcal{F} \cup \{0\}, L_2(\mathbb{P}_n))}\, d\epsilon, \tag{41}$$

where $\mathbb{E}_\varepsilon$ is the expectation with respect to the Rademacher variables, given fixed $X_1, \ldots, X_n$ (note that $\log N(\epsilon, \mathcal{F} \cup \{0\}, L_2(\mathbb{P}_n)) = 0$ for any $\epsilon > \sigma_{n,2}$). Making a change of variable and bounding the random entropy by a supremum we see that the right side is bounded by

$$\int_0^{\sigma_{n,2}/\|F\|_n} \sqrt{\log N(\epsilon \|F\|_n, \mathcal{F} \cup \{0\}, L_2(\mathbb{P}_n))}\, d\epsilon \|F\|_n \leq J(\theta_n, \mathcal{F}, F)\|F\|_n.$$

Next, by taking the expectation over $X_1, \ldots, X_n$ we obtain the first inequality of the theorem.

Since $\theta_n \leq 1$, we have that $J(\theta_n, \mathcal{F}, F) \leq J(1, \mathcal{F}, F)$. Furthermore, by Jensen's inequality applied to the root function, $\mathbb{E}\|F\|_n \leq \sqrt{\mathbb{E}[n^{-1} \sum_{i=1}^n F^2(X_i)]} = \|F\|_{P,2}$. This gives the inequality on the right side of the theorem. $\square$

The above theorem shows that the order of magnitude of $\|\mathbb{G}_n\|_{\mathcal{F}}$ is not bigger than $J(1, \mathcal{F}, F)$ times the order of $\|F\|_{P,2}$, which is the order of magnitude of the random variable $|\mathbb{G}_n(F)|$ if the entropy integral is finite.

Example 4.8 (Supremum of the empirical process). *Recall the setting of Section 3.7. Suppose that $\mathcal{F}$ is a class of B-uniformly bounded functions such that*

$$N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n}) \le C\nu(16e)^{\nu}\left(\frac{B}{\epsilon}\right)^{2\nu}.$$

We will see in Section 7 that if $\mathcal{F}$ is a function class with finite VC dimension ν, then the above inequality holds. The goal is to study the expected supremum of the empirical process over $\mathcal{F}$, i.e., $\|\mathbb{P}_n - P\|_{\mathcal{F}}$. From our previous results, we have seen that by exploiting concentration (see Section 3.7) and symmetrization results (see Theorem 3.17), the study of $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ can be reduced to controlling the expectation $\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$. We consider the random variable $Z_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)$, and consider the stochastic process $\{Z_f : f \in \mathcal{F}\}$. We have seen that by Lemma 3.12, the increment $Z_f - Z_g$ is sub-Gaussian with parameter $\|f-g\|_n^2$. Consequently, by Dudley's entropy integral (see (41)), we have

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] &\le \frac{24}{\sqrt{n}}\int_0^B \sqrt{\log N(\epsilon, \mathcal{F}\cup\{0\}, \|\cdot\|_{\mathbb{P}_n})}\, d\epsilon, \\ &\le c_0\frac{1}{\sqrt{n}}\int_0^B \sqrt{\log[1 + c(B/\epsilon)^{2\nu}]}\, d\epsilon = c_0' B\sqrt{\frac{\nu}{n}},\end{aligned}$$

since the integral is finite[43]; here c, c_0 and c_0' are constants. Thus,

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \lesssim B\sqrt{\frac{\nu}{n}}.$$

Example 4.9. *Suppose that $\mathcal{F} = \{m_\theta(\cdot) : \theta \in \Theta\}$ is a parameterized class such that $\mathcal{F} = -\mathcal{F}$, where $\Theta = B(0;1) \subset \mathbb{R}^d$ is the unit Euclidean ball in $\mathbb{R}^d$. Suppose that the class of functions is L-Lipschitz with respect to the Euclidean distance on $\mathbb{R}^d$ so that for all x,*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le L\|\theta_1 - \theta_2\|.$$

Exercise (HW2): Show that $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} = O\left(L\sqrt{\frac{d}{n}}\right)$.

Observe that the right side of (40) depends on the $L_2(P)$-norm of the envelope function F, which may, in some situations, be large compared with the maximum $L_2(P)$-norm of functions in $\mathcal{F}$, namely, $\sigma := \sup_{f\in\mathcal{F}}\|f\|_{P,2}$. In such a case, the following theorem will be more useful.

Theorem 4.10 ([van der Vaart and Wellner, 2011], [Chernozhukov et al., 2014][44]). *Suppose that $0 < \|F\|_{P,2} < \infty$, and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} Pf^2 \le$*

[43] Note that

$$\int_0^B \sqrt{\log[1 + c(B/\epsilon)^{2\nu}]}\, d\epsilon = B\int_0^1 \sqrt{\log[1 + c(1/\epsilon)^{2\nu}]}\, d\epsilon \lesssim B\sqrt{\nu}\int_0^1 \sqrt{\log(1/\epsilon)}\, d\epsilon.$$

[44] A version of this result was proved in [van der Vaart and Wellner, 2011] under the additional assumption that the envelope F is bounded; the current version is due to [Chernozhukov et al., 2014]. We will skip the proof of this result.

$\sigma^2 \le \|F\|_{P,2}^2$. *Let* $\delta := \sigma/\|F\|_{P,2}$. *Define* $B := \sqrt{\mathbb{E}[\max_{1\le i\le n} F^2(X_i)]}$. *Then,*

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim J(\delta, \mathcal{F}, F)\|F\|_{P,2} + \frac{BJ^2(\delta, \mathcal{F}, F)}{\delta^2\sqrt{n}}. \tag{42}$$

4.3 Maximal inequalities with bracketing

As one might expect, there exists maximal inequalities that work with bracketing numbers (as opposed to covering numbers). However, the bracketing result is more delicate and difficult to prove[45]. We will just state the result and illustrate an application of the result.

Recall that, apart from the constant $1/2$, bracketing numbers are bigger than covering numbers. The advantage of a bracket is that it gives pointwise control over a function: $l(x) \le f(x) \le u(x)$, for every $x \in \mathcal{X}$. The maximal inequalities in the preceding subsection (without bracketing) compensate this lack of pointwise control by considering entropy under every measure Q, not just the law P of the observations. With bracketing we can obtain analogous results using only bracketing numbers under P.

Definition 4.11 (Bracketing integral)**.** *The bracketing integral is defined as*

$$J_{[\,]}(\delta, \mathcal{F}, L_2(P)) := \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}\cup\{0\}, L_2(P))}\, d\epsilon < \infty, \qquad \delta > 0.$$

Theorem 4.12. *For a class $\mathcal{F}$ of measurable functions with envelope F,*

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{F}}\big] \lesssim J_{[\,]}(\|F\|_{P,2}, \mathcal{F}\cup\{0\}, L_2(P)).$$

The preceding theorem does not take the size of the functions f into account. The following theorem remedy this, which is however restricted to uniformly bounded classes.

Theorem 4.13 (Lemma 3.4.2 of [van der Vaart and Wellner, 1996])**.** *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $Pf^2 < \delta^2$ and $\|f\|_\infty \le M$ for every f,*

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim J_{[\,]}(\delta, \mathcal{F}, L_2(P))\left(1 + \frac{J_{[\,]}(\delta, \mathcal{F}, L_2(P))}{\delta^2\sqrt{n}}M\right).$$

4.4 Bracketing number for some function classes

Here are some important results about the bracketing numbers of nonparametric classes of functions. A good account on this is [van der Vaart and Wellner, 1996, Section 2.7].

1. Let $C_1^\alpha(E)$, for a bounded convex subset E of $\mathbb{R}^d$ with nonempty interior, be the set of functions $f : E \to \mathbb{R}$ with $\|f\|_\infty \le 1$ and with degree of smoothness α (if $\alpha \le 1$, Hölder of order α and constant 1, and if $\alpha > 1$, differentiable up to order $\lfloor\alpha\rfloor$, the

[45] In this subsection we just state a few results without proofs; see [van der Vaart and Wellner, 1996, Chapter 2.14] for a more detailed discussion with complete proofs of these results.

greatest integer smaller than α, with all the partial derivatives of order α, Hölder of order $\alpha - \lfloor \alpha \rfloor$ and constant 1, and with all the partial derivatives bounded by 1. Then,

$$\log N_{[\,]}(\epsilon, C_1^\alpha(E), L_r(Q)) \leq K(1/\epsilon)^{d/\alpha}$$

for all $r \geq 1$, $\epsilon > 0$ and probability measure Q on $\mathbb{R}^d$, where K is a constant that depends only on α, $\operatorname{diam}(E)$ and d.

2. The class $\mathcal{F}$ of monotone functions $\mathbb{R}$ to $[0, 1]$ satisfies

$$\log N_{[\,]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K/\epsilon,$$

for all Q and all $r \geq 1$, for K depending only on r.

5 Rates of convergence of M-estimators

Let (Θ, d) be a semimetric space. As usual, we are given i.i.d. observations $X_1, X_2, \ldots, X_n$ from a probability distribution P on $\mathcal{X}$. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ denote a stochastic process and let $\{M(\theta) : \theta \in \Theta\}$ denote a deterministic process. Suppose $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\theta)$ and suppose θ_0 maximizes $M(\theta)$, i.e.,

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \mathbb{M}_n(\theta), \qquad \text{and} \qquad \theta_0 = \operatorname*{argmax}_{\theta \in \Theta} M(\theta).$$

We want to find the rate δ_n of the convergence of $\hat{\theta}_n$ to θ_0 in the metric d, i.e., $d(\hat{\theta}_n, \theta_0)$. A *rate of convergence*[46] of δ_n means that

$$\delta_n^{-1} d(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}(1).$$

We assume that $\mathbb{M}_n(\theta)$ gets close to $M(\theta)$ as n increases and under this setting want to know how close $\hat{\theta}_n$ is to θ_0.

5.1 The rate theorem

If the metric d is chosen appropriately we may expect that the asymptotic criterion decreases *quadratically*[47] when θ moves away from θ_0:

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0) \tag{43}$$

for all $\theta \in \Theta$.

Consider the probability $\mathbb{P}\big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\big)$ for a large M. We want to understand for which δ_n this probability becomes small as M grows large. Write

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) = \sum_{j > M} \mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big).$$

Let us define the "shells" $S_j := \big\{\theta \in \Theta : 2^{j-1}\delta_n < d(\theta, \theta_0) \le 2^j \delta_n\big\}$ so that

$$\mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big) = \mathbb{P}\Big(\hat{\theta}_n \in S_j\Big).$$

[46]Recall that a sequence of random variables $\{Z_n\}$ is said to be *bounded in probability* or $O_{\mathbb{P}}(1)$ if

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}(|Z_n| > T) = 0.$$

In other words, $Z_n = O_{\mathbb{P}}(1)$, if for any given $\epsilon > 0$, $\exists\, T_\epsilon, N_\epsilon > 0$ such that $\mathbb{P}(|Z_n| > T_\epsilon) < \epsilon$ for all $n \ge N_\epsilon$.

[47]To get intuition about this condition assume that if $M : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and $d(\cdot, \cdot)$ is the Euclidean distance, then, for θ in a neighborhood of θ_0,

$$M(\theta) - M(\theta_0) = \nabla M(\theta_0)^\top (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top \nabla^2 M(\tilde{\theta}_0)(\theta - \theta_0) \le -c\|\theta - \theta_0\|^2$$

where $\nabla M(\theta_0) = 0$ (as θ_0 is a maximizer of $M(\cdot)$) and $\nabla^2 M(\theta_0)$ (the Hessian matrix of $M(\cdot)$) is assumed to be negative definite, in which case we can find such a constant $c > 0$ (corresponding to the smallest eigenvalue of $\nabla^2 M(\theta_0)$); here $\tilde{\theta}_0$ is a point close to θ_0.

As $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\theta)$, it is obvious that

$$\mathbb{P}\Big(\hat{\theta}_n \in S_j\Big) \le \mathbb{P}\Big(\sup_{\theta \in S_j} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \ge 0\Big).$$

Now $d(\theta, \theta_0) > 2^{j-1}\delta_n$ for $\theta \in S_j$ which implies, by (43), that

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0) \lesssim -2^{2j-2}\delta_n^2 \qquad \text{for } \theta \in S_j \tag{44}$$

or $\sup_{\theta \in S_j} [M(\theta) - M(\theta_0)] \lesssim -2^{2j-2}\delta_n^2$. Thus, the event $\sup_{\theta \in S_j} [\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)] \ge 0$ can only happen if $\mathbb{M}_n$ and M are not too close. Let

$$U_n(\theta) := \mathbb{M}_n(\theta) - M(\theta), \qquad \text{for } \theta \in \Theta.$$

It follows from (44) that

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\theta \in S_j} [\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)] \ge 0\Big) &\le \mathbb{P}\Big(\sup_{\theta \in S_j} [U_n(\theta) - U_n(\theta_0)] \gtrsim 2^{2j-2}\delta_n^2\Big) \\
&\le \mathbb{P}\left(\sup_{\theta : d(\theta,\theta_0) \le 2^j \delta_n} [U_n(\theta) - U_n(\theta_0)] \gtrsim 2^{2j-2}\delta_n^2\right) \\
&\lesssim \frac{1}{2^{2j-2}\delta_n^2} \mathbb{E}\left[\sup_{\theta : d(\theta,\theta_0) \le 2^j \delta_n} (U_n(\theta) - U_n(\theta_0))\right].
\end{aligned}$$

Suppose that there is a function $\phi_n(\cdot)$ such that

$$\mathbb{E}\left[\sup_{\theta : d(\theta,\theta_0) \le u} \sqrt{n}(U_n(\theta) - U_n(\theta_0))\right] \lesssim \phi_n(u) \qquad \text{for every } u > 0. \tag{45}$$

We thus get

$$\mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big) \lesssim \frac{\phi_n(2^j \delta_n)}{\sqrt{n} 2^{2j} \delta_n^2}$$

for every j. As a consequence,

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) \lesssim \frac{1}{\sqrt{n}} \sum_{j > M} \frac{\phi_n(2^j \delta_n)}{2^{2j} \delta_n^2}.$$

The following assumption on $\phi_n(\cdot)$ is usually made to simplify the expression above: there exists $0 < \alpha < 2$ such that

$$\phi_n(cx) \le c^\alpha \phi_n(x) \qquad \text{for all } c > 1 \text{ and } x > 0. \tag{46}$$

Under this assumption, we get

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) \lesssim \frac{\phi_n(\delta_n)}{\sqrt{n}\delta_n^2} \sum_{j > M} 2^{j(\alpha - 2)}.$$

The quantity $\sum_{j > M} 2^{j(\alpha-2)}$ converges to zero as $M \to \infty$. Observe that if we further assume that

$$\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2, \qquad \text{as } n \text{ varies,} \tag{47}$$

then

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) \leq c \sum_{j>M} 2^{j(\alpha-2)},$$

for a constant $c > 0$ (which does not depend on n, M). Let u_M denote the right side of the last display. It follows therefore that, under assumptions (46) and (106), we get

$$d(\hat{\theta}_n, \theta_0) \leq 2^M \delta_n \qquad \text{with probability at least } 1 - u_M, \quad \text{for all } n.$$

Further note that $u_M \to 0$ as $M \to \infty$. This gives us the following non-asymptotic rate of convergence theorem.

Theorem 5.1. *Let (Θ, d) be a semi-metric space. Fix $n \geq 1$. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a stochastic process and $\{M(\theta) : \theta \in \Theta\}$ be a deterministic process. Assume condition (43) and that the function $\phi_n(\cdot)$ satisfies (45) and (46). Then for every $M > 0$, we get $d(\hat{\theta}_n, \theta_0) \leq 2^M \delta_n$ with probability at least $1 - u_M$ provided (106) holds. Here $u_M \to 0$ as $M \to \infty$.*

Suppose now that condition (43) holds only for θ in a neighborhood of θ_0 and that (45) holds only for small u. Then one can prove the following asymptotic result under the additional condition that $\hat{\theta}_n$ is consistent (i.e., $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathbb{P}} 0$).

Theorem 5.2 (Rate theorem)**.** *Let Θ be a semi-metric space. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a stochastic process and $\{M(\theta) : \theta \in \Theta\}$ be a deterministic process. Assume that (43) is satisfied for every θ in a neighborhood of θ_0. Also, assume that for every n and sufficiently small u condition (45) holds for some function ϕ_n satisfying (46), and that (106) holds. If the sequence $\hat{\theta}_n$ satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_{\mathbb{P}}(\delta_n^2)$ and if $\hat{\theta}_n$ is consistent in estimating θ_0, then $d(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}(\delta_n)$.*

Proof. The above result is Theorem 3.2.5 in [van der Vaart and Wellner, 1996] where you can find its proof. The proof is very similar to the proof of Theorem 5.1. The crucial observation is to realize that: for any $\eta > 0$,

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) \leq \sum_{j>M, 2^{j-1}\delta_n \leq \eta} \mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n\Big) + \mathbb{P}\Big(2d(\hat{\theta}_n, \theta_0) > \eta\Big).$$

The first term can be tackled as before while the second term goes to zero by the consistency of $\hat{\theta}_n$. □

Remark 5.1. *In the case of i.i.d. data and criterion functions of the form $\mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta]$ and $M(\theta) = P[m_\theta]$, the centered and scaled process $\sqrt{n}(\mathbb{M}_n - M)(\theta) = \mathbb{G}_n[m_\theta]$ equals the empirical process at m_θ. Condition (45) involves the suprema of the empirical process indexed by classes of functions*

$$\mathcal{M}_u := \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) \leq u\}.$$

Thus, we need to find the existence of $\phi_n(\cdot)$ such that $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{M}_u} \lesssim \phi_n(u)$.

Remark 5.2. *Theorem 5.2 gives the correct rate in fair generality, the main problem being to derive sharp bounds on the modulus of continuity of the empirical process. A simple, but not necessarily efficient, method is to apply the maximal inequalities (with and without bracketing). These yield bounds in terms of the uniform entropy integral* $J(1, \mathcal{M}_u, M_u)$ *or the bracketing integral* $J_{[\,]}(\|M_u\|_{P,2}, \mathcal{M}_u, L_2(P))$ *of the class* $\mathcal{M}_u$ *given by*

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{M}_u}\big] \lesssim J(1, \mathcal{M}_u, M_u)[P(M_u^2)]^{1/2} \tag{48}$$

where

$$J(1, \mathcal{M}_u, M_u) = \int_0^1 \sup_Q \sqrt{\log N(\epsilon \|M_u\|_{Q,2}, \mathcal{M}_u, L_2(Q))}\, d\epsilon$$

and

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{M}_u}\big] \lesssim J_{[\,]}(\|M_u\|, \mathcal{M}_u, L_2(P)),$$

where

$$J_{[\,]}(\delta, \mathcal{M}_u, L_2(P)) = \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{M}_u, L_2(P))}\, d\epsilon.$$

Here M_u *is the envelope function of the class* $\mathcal{M}_u$*. In this case, we can take* $\phi_n^2(u) = P[M_u^2]$ *and this leads to a rate of convergence* δ_n *of at least the solution of*

$$P[M_{\delta_n}^2] \sim n\delta_n^4.$$

Observe that the rate of convergence in this case is driven by the sizes of the envelope functions as $u \downarrow 0$*, and the size of the classes is important only to guarantee a finite entropy integral.*

Remark 5.3. *In genuinely infinite-dimensional situations, this approach could be less useful, as it is intuitively clear that the precise entropy must make a difference for the rate of convergence. In this situation, the maximal inequalities obtained in Section 4 may be used.*

Remark 5.4. *For a Euclidean parameter space, the first condition of the theorem is satisfied if the map* $\theta \mapsto Pm_\theta$ *is twice continuously differentiable at the point of maximum* θ_0 *with a nonsingular second-derivative matrix.*

5.2 Some examples

5.2.1 Euclidean parameter

Let $X_1, \ldots, X_n$ be i.i.d. random elements on $\mathcal{X}$ with a common law P, and let $\{m_\theta : \theta \in \Theta\}$ be a class of real-valued measurable maps. Suppose that $\Theta \subset \mathbb{R}^d$, and that, for every $\theta_1, \theta_2 \in \Theta$ (or just in a neighborhood of θ_0),

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le F(x)\|\theta_1 - \theta_2\| \tag{49}$$

for some measurable function $F : \mathcal{X} \to \mathbb{R}$ with $PF^2 < \infty$. Then the class of functions $\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| \le \delta\}$ has envelope function δF and bracketing number (see Theorem 2.14) satisfying

$$N_{[\,]}(2\epsilon\|F\|_{P,2}, \mathcal{M}_\delta, L_2(P)) \le N(\epsilon, \{\theta : \|\theta - \theta_0\| \le \delta\}, \|\cdot\|) \le \left(\frac{C\delta}{\epsilon}\right)^d,$$

where the last inequality follows from Lemma 2.7 coupled with the fact that the ϵ-covering number of δB (for any set B) is the ϵ/δ-covering number of B. In view of the maximal inequality with bracketing (see Theorem 4.12),

$$\mathbb{E}_P\big[\|\mathbb{G}_n\|_{\mathcal{M}_\delta}\big] \lesssim \int_0^{\delta\|F\|_{P,2}} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{M}_\delta, L_2(P))}\, d\epsilon \lesssim \delta\|F\|_{P,2}.$$

Thus, we can take $\phi_n(\delta) \asymp \delta$, and the inequality $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ is solved by $\delta_n = 1/\sqrt{n}$. We conclude that the rate of convergence of $\hat{\theta}_n$ is $n^{-1/2}$ as soon as $P(m_\theta - m_{\theta_0}) \le -c\|\theta - \theta_0\|^2$, for every $\theta \in \Theta$ in a neighborhood of θ_0.

Example 5.3 (Least absolute deviation regression)**.** *Given i.i.d. random vectors $Z_1, \ldots, Z_n$, and $e_1, \ldots, e_n$ in $\mathbb{R}^d$ and $\mathbb{R}$, respectively, let*

$$Y_i = \theta_0^\top Z_i + e_i.$$

The least absolute-deviation estimator $\hat{\theta}_n$ minimizes the function

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^n |Y_i - \theta^\top Z_i| = \mathbb{P}_n m_\theta,$$

where $\mathbb{P}_n$ is the empirical measure of $X_i := (Z_i, Y_i)$, and $m_\theta(x) = |y - \theta^\top z|$.

Exercise (HW2): Show that the parameter θ_0 is a point of minimum of the map $\theta \mapsto P|Y - \theta^\top Z|$ if the distribution of the error e_1 has median zero. Furthermore, show that the maps $\theta \mapsto m_\theta$ satisfies condition (49):

$$\Big||y - \theta_1^\top z| - |y - \theta_2^\top z|\Big| \le \|\theta_1 - \theta_2\|\|z\|.$$

Argue the consistency of the least-absolute-deviation estimator from the convexity of the map $\theta \mapsto |y - \theta^\top z|$. Moreover, show that the map $\theta \mapsto P|Y - \theta^\top Z|$ is twice differentiable at θ_0 if the distribution of the errors has a positive density at its median (you may need to assume that Z and e are integrable and $\mathbb{E}[ZZ^\top]$ is positive definite). Furthermore, derive the rate of convergence of $\hat{\theta}_n$ in this situation.

5.2.2 A non-standard example

Example 5.4 (Analysis of the shorth)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with a differentiable density p with respect to the Lebesgue measure. Let F_X be the distribution* *function of X. Suppose that p is a unimodal (bounded) continuously differentiable symmetric density with mode θ_0 (with $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$). We want to estimate θ_0.*

Exercise (HW2): Let

$$\mathbb{M}(\theta) := Pm_\theta = \mathbb{P}(|X - \theta| \le 1) = F_X(\theta + 1) - F_X(\theta - 1)$$

where $m_\theta(x) = \mathbf{1}_{[\theta-1,\theta+1]}(x)$. Show that $\theta_0 = \mathrm{argmax}_{\theta\in\mathbb{R}} \mathbb{M}(\theta)$. Thus, θ_0 is the center of an interval of length 2 that contains the largest possible (population) fraction of data points. We can estimate θ_0 by

$$\hat{\theta}_n := \underset{\theta\in\mathbb{R}}{\mathrm{argmax}}\, \mathbb{M}_n(\theta), \qquad \textit{where} \qquad \mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta].$$

Show that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$? The functions $m_\theta(x) = \mathbf{1}_{[\theta-1,\theta+1]}(x)$ are not Lipschitz in the parameter $\theta \in \Theta \equiv \mathbb{R}$. Nevertheless, the classes of functions $\mathcal{M}_\delta$ satisfy the conditions of Theorem 5.2. These classes have envelope function

$$\sup_{|\theta-\theta_0|\le\delta} \left| \mathbf{1}_{[\theta-1,\theta+1]} - \mathbf{1}_{[\theta_0-1,\theta_0+1]} \right| \le \mathbf{1}_{[\theta_0-1-\delta,\theta_0-1+\delta]} + \mathbf{1}_{[\theta_0+1-\delta,\theta_0+1+\delta]}.$$

The $L_2(P)$-norm of these functions is bounded above by a constant times $\sqrt{\delta}$. Thus, the conditions of the rate theorem are satisfied with $\phi_n(\delta) = c\sqrt{\delta}$ for some constant c, leading to a rate of convergence of $n^{-1/3}$. We will show later that $n^{1/3}(\hat{\theta}_n - \theta_0)$ converges in distribution to a non-normal limit as $n \to \infty$.

Example 5.5 (A toy change point problem)**.** *Suppose that we have i.i.d. data $\{X_i = (Z_i, Y_i) : i = 1, \ldots, n\}$ where $Z_i \sim \mathrm{Unif}(0, 1)$ and*

$$Y_i = \mathbf{1}_{[0,\theta_0]}(Z_i) + \epsilon_i, \qquad\qquad \textit{for } i = 1, \ldots, n.$$

Here, ϵ_i's are the unobserved errors assumed to be i.i.d. $N(0, \sigma^2)$. Further, for simplicity, we assume that ϵ_i is independent of Z_i. The goal is to estimate the unknown parameter $\theta_0 \in (0, 1)$. A natural procedure is to consider the least squares estimator:

$$\hat{\theta}_n := \underset{\theta\in[0,1]}{\mathrm{argmin}}\, \mathbb{P}_n[(Y - \mathbf{1}_{[0,\theta]}(Z))^2].$$

Exercise (HW2): Show that $\hat{\theta}_n := \mathrm{argmax}_{\theta\in[0,1]} \mathbb{M}_n(\theta)$ where

$$\mathbb{M}_n(\theta) := \mathbb{P}_n[(Y - 1/2)\{\mathbf{1}_{[0,\theta]}(Z) - \mathbf{1}_{[0,\theta_0]}(Z)\}].$$

Prove that $\mathbb{M}_n$ converges uniformly to

$$M(\theta) := P[(Y - 1/2)\{\mathbf{1}_{[0,\theta]}(Z) - \mathbf{1}_{[0,\theta_0]}(Z)\}].$$

Show that $M(\theta) = -|\theta - \theta_0|/2$. As a consequence, show that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

To find the rate of convergence of $\hat{\theta}_n$ we consider the metric $d(\theta_1, \theta_2) := \sqrt{|\theta_1 - \theta_2|}$. Show that the conditions needed to apply Theorem 5.2 hold with this choice of $d(\cdot, \cdot)$. Using Theorem 5.2 derive that $n(\hat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$.

5.2.3 Persistency in high-dimensional regression

Let $Z^i := (Y^i, X^i_1, \ldots, X^i_p)$, $i = 1, \ldots, n$, be i.i.d. random vectors, where $Z^i \sim P$. It is desired to predict Y by $\sum_j \beta_j X_j$, where $(\beta_1, \ldots, \beta_p) \in B_n \subset \mathbb{R}^p$, under a prediction loss. We assume that $p = n^\alpha$, $\alpha > 0$, that is, there could be many more explanatory variables than observations. We consider sets B_n restricted by the maximal number of non-zero coefficients of their members, or by their l_1-radius. We study the following asymptotic question: how 'large' may the set B_n be, so that it is still possible to select empirically a predictor whose risk under P is close to that of the best predictor in the set?

We formulate this problem using a triangular array setup, i.e., we model the observations $Z^1_n, \ldots, Z^n_n$ as i.i.d. random vectors in $\mathbb{R}^{p_n+1}$, having distribution P_n (that depends on n). In the following we will hide the dependence on n and just write $Z^1, \ldots, Z^n$. We will consider B_n of the form

$$B_{n,b} := \{\beta \in \mathbb{R}^{p_n} : \|\beta\|_1 \leq b\}, \tag{50}$$

where $\|\cdot\|_1$ denotes the l_1-norm. For any $Z := (Y, X_1, \ldots, X_p) \sim P$, we will denote the expected prediction error by

$$L_P(\beta) := \mathbb{E}_P\Big[(Y - \sum_{j=1}^{p} \beta_j X_j)^2\Big] = \mathbb{E}_P\Big[(Y - \beta^\top X)^2\Big]$$

where $X = (X_1, \ldots, X_p)$. The best linear predictor, where $Z \sim P_n$, is given by

$$\beta^*_n := \arg\min_{\beta \in B_{n,b_n}} L_{P_n}(\beta),$$

for some sequence of $\{b_n\}_{n\geq 1}$. We estimate the best linear predictor β^*_n from the sample by

$$\hat\beta_n := \arg\min_{\beta \in B_{n,b_n}} L_{\mathbb{P}_n}(\beta) = \arg\min_{\beta \in B_{n,b_n}} \frac{1}{n}\sum_{i=1}^{n}(Y^i - \beta^\top X^i)^2,$$

where $\mathbb{P}_n$ is the empirical measure of the Z^i's.

We study the following asymptotic question: how 'large' may the set B_{n,b_n} be, so that it is still possible to select empirically a predictor whose risk under P_n is close to that of the best predictor in the set?

We say that $\hat\beta_n$ is *persistent* (relative to B_{n,b_n} and P_n) ([Greenshtein and Ritov, 2004]) if and only if

$$L_{P_n}(\hat\beta_n) - L_{P_n}(\beta^*_n) \xrightarrow{\mathbb{P}} 0.$$

This is certainly a weak notion of "risk-consistency" — we are only trying to consistently estimate the expected predictor error. However, as we will see soon, this notion does not require any modeling assumptions on the (joint) distribution of Z (in particular, we are not assuming that there is a 'true' linear model). The following theorem is a version of Theorem 3 in [Greenshtein and Ritov, 2004].

Theorem 5.6. *Suppose that $p_n = n^\alpha$, where $\alpha > 0$. Let*

$$F(Z^i) := \max_{0\le j,k\le p} |X_j^i X_k^i - \mathbb{E}_{P_n}(X_j^i X_k^i)|, \quad \text{where we take } X_0^i = Y^i, \text{ for } i = 1, \ldots, n.$$

Suppose that $\mathbb{E}_{P_n}[F^2(Z^1)] \le M < \infty$, for all n. Then for $b_n = o((n/\log n)^{1/4})$, $\hat{\beta}_n$ is persistent relative to B_{n,b_n}.

Proof. From the definition of β_n^* and $\hat{\beta}_n$ it follows that

$$L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \ge 0, \qquad \text{and} \qquad L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \le 0.$$

Thus,

$$\begin{aligned} 0 &\le L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \\ &= \Big(L_{P_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\hat{\beta}_n)\Big) + \Big(L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*)\Big) + \Big(L_{\mathbb{P}_n}(\beta_n^*) - L_{P_n}(\beta_n^*)\Big) \\ &\le 2 \sup_{\beta\in B_{n,b_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)|, \end{aligned}$$

where we have used the fact that $L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \le 0$. To simply our notation, let $\gamma = (-1, \beta) \in \mathbb{R}^{p_n+1}$. Then $L_{P_n}(\beta) = \gamma^\top \Sigma_{P_n} \gamma$ and $L_{\mathbb{P}_n}(\beta) = \gamma^\top \Sigma_{\mathbb{P}_n} \gamma$ where $\Sigma_{P_n} = \Big(\mathbb{E}_{P_n}(X_j^1 X_k^1)\Big)_{0\le j,k\le p_n}$ and $\Sigma_{\mathbb{P}_n} = \Big(\frac{1}{n}\sum_{i=1}^n X_j^i X_k^i\Big)_{0\le j,k\le p_n}$. Thus,

$$|L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| \le |\gamma^\top(\Sigma_{\mathbb{P}_n} - \Sigma_{P_n})\gamma| \le \|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty \|\gamma\|_1^2,$$

where $\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = \sup_{0\le j,k\le p_n} \left|\frac{1}{n}\sum_{i=1}^n X_j^i X_k^i - E_{P_n}(X_j^1 X_k^1)\right|$. Therefore,

$$\begin{aligned} \mathbb{P}\big(L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) > \epsilon\big) &\le \mathbb{P}\Big(2 \sup_{\beta\in B_{n,b_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| > \epsilon\Big) \\ &\le \mathbb{P}\Big(2(b_n+1)^2 \|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty > \epsilon\Big) \\ &\le \frac{2(b_n+1)^2}{\epsilon} \mathbb{E}\Big[\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty\Big]. \end{aligned} \tag{51}$$

Let $\mathcal{F} = \{f_{j,k} : 0 \le j, k \le p_n\}$ where $f_{j,k}(z) := x_j x_k - \mathbb{E}_{P_n}(X_j^1 X_k^1)$ and $z = (x_0, x_1, \ldots, x_{p_n})$. Observe that $\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = \|\mathbb{P}_n - P_n\|_\mathcal{F}$. We will now use the following maximal inequality with bracketing entropy (see Theorem 4.12):

$$\mathbb{E}\|\sqrt{n}(\mathbb{P}_n - P)\|_\mathcal{F} \lesssim J_{[\,]}(\|F_n\|_{P_n,2}, \mathcal{F} \cup \{0\}, L_2(P_n)),$$

where F_n is an envelope of $\mathcal{F}$. Note that F_n can be taken as F (defined in the statement of the theorem). We can obviously cover $\mathcal{F}$ with the ϵ-brackets $[f_{j,k} - \epsilon/2, f_{j,k} + \epsilon/2]$, for every $\epsilon > 0$, and thus, $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n)) \le (p_n+1)^2$. Therefore, using (51) and the maximal inequality above,

$$\mathbb{P}\big(L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) > \epsilon\big) \lesssim \frac{2(b_n+1)^2}{\epsilon} \frac{\sqrt{2\log(p_n+1)}}{\sqrt{n}} \sqrt{M} \lesssim \frac{b_n^2\sqrt{\alpha \log n}}{\sqrt{n}} \to 0,$$

as $n \to \infty$, by the assumption on b_n. ∎

6 Rates of convergence of infinite dimensional parameters

If Θ is an infinite-dimensional set, such as a function space, then maximization of a criterion over the full space may not always be a good idea. For instance, consider fitting a function $\theta : [0,1] \to \mathbb{R}$ to a set of observations $(z_1, Y_1), \dots, (z_n, Y_n)$ by least squares, i.e., we minimize

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \theta(z_i)\}^2.$$

If Θ consists of all functions $\theta : [0,1] \to \mathbb{R}$, then obviously the minimum is 0, taken for any function that interpolates the data points exactly: $\theta(z_i) = Y_i$ for every $i = 1, \dots, n$. This interpolation is typically not a good estimator, but *overfits* the data: it follows the given data exactly even though these probably contain error. The interpolation very likely gives a poor representation of the true regression function.

One way to rectify this problem is to consider minimization over a restricted class of functions. For example, the minimization can be carried out over all functions with 2 derivatives, which are bounded above by 10 throughout the interval; here the numbers 2 and (particularly) 10 are quite arbitrary. To prevent overfitting the size of the derivatives should not be too large, but can grow as we obtain more samples.

The *method of sieves* is an attempt to implement this. Sieves are subsets $\Theta_n \subset \Theta$, typically increasing in n, that can approximate any given function θ_0 that is considered likely to be "true". Given n observations the maximization is restricted to Θ_n, and as n increases this "sieve" is taken larger. In this section we extend the rate theorem in the previous section to sieved M-estimators, which include maximum likelihood estimators and least-squares estimators.

We also generalize the notation and other assumptions. In the next theorem the empirical criterion $\theta \mapsto \mathbb{P}_n m_\theta$ is replaced by a general stochastic process

$$\theta \mapsto \mathbb{M}_n(\theta).$$

It is then understood that each "estimator" $\hat{\theta}_n$ is a map defined on the same probability space as $\mathbb{M}_n$, with values in the index set Θ_n (which may be arbitrary set) of the process $\mathbb{M}_n$.

Corresponding to the criterion functions are *centering functions* $\theta \mapsto M_n(\theta)$ and "true parameters" $\theta_{n,0}$. These may be the mean functions of the processes $\mathbb{M}_n$ and their point of maximum, but this is not an assumption.

In this generality we also need not assume that Θ_n is a metric space, but measure the "discrepancy" or "distance" between θ and the true "value" $\theta_{n,0}$ by a map $\theta \mapsto d_n(\theta, \theta_{n,0})$ from Θ_n to $[0, \infty)$.

Theorem 6.1 (Rate of convergence). *For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set $\Theta_n \cup \{\theta_{n,0}\}$, and let $\theta \mapsto d_n(\theta, \theta_{n,0})$ be an arbitrary map from Θ_n to $[0, \infty)$.*

Let $\tilde{\delta}_n \geq 0$ and suppose that, for every n and $\delta > \tilde{\delta}_n$,

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} [M_n(\theta) - M_n(\theta_{n,0})] \leq -c\delta^2, \tag{52}$$

for some $c > 0$ (for all $n \geq 1$) and

$$\mathbb{E}\left[\sup_{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta} \sqrt{n}\Big|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})\Big|\right] \lesssim \phi_n(\delta),$$

for increasing functions $\phi_n : [\tilde{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $0 < \alpha < 2$. Let $\theta_n \in \Theta_n$ and let δ_n satisfy

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \qquad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \qquad \delta_n \geq \tilde{\delta}_n.$$

If the sequence $\hat{\theta}_n$ takes values in Θ_n and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_{\mathbb{P}}(\delta_n^2)$, then

$$d_n(\hat{\theta}_n, \theta_{n,0}) = O_{\mathbb{P}}(\delta_n).$$

Exercise (HW2): Complete the proof. Hint: The proof is similar to that of the previous rate theorem. That all entities are now allowed to depend on n asks for notational changes only, but the possible discrepancy between θ_n and $\theta_{n,0}$ requires some care.

The theorem can be applied with $\hat{\theta}_n$ and $\theta_{n,0}$ equal to the maximizers of $\theta \mapsto \mathbb{M}_n(\theta)$ over a sieve Θ_n and of $\theta \mapsto M_n(\theta)$ over a full parameter set Θ, respectively. Then (52) requires that the centering functions fall off quadratically in the "distance" $d_n(\theta, \theta_{n,0})$ as θ moves away from the maximizing value $\theta_{n,0}$. We use $\tilde{\delta}_n = 0$, and the theorem shows that the "distance" of $\hat{\theta}_n$ to $\theta_{n,0}$ satisfies

$$d_n^2(\hat{\theta}_n, \theta_{n,0}) = O_{\mathbb{P}}(\delta_n^2 + M_n(\theta_{n,0}) - M_n(\theta_n)), \tag{53}$$

for δ_n solving $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ and for any $\theta_n \in \Theta_n$. Thus the rate δ_n is determined by the "modulus of continuity" $\delta \mapsto \phi_n(\delta)$ of the centered processes $\sqrt{n}(\mathbb{M}_n - M_n)$ over Θ_n and the discrepancy $M_n(\theta_{n,0}) - M_n(\theta_n)$. The latter vanishes if $\theta_n = \theta_{n,0}$ but this choice of θ_n may not be admissible (as θ_n must be an element of the sieve and $\theta_{n,0}$ need not). A natural choice of θ_n is to take θ_n as the closest element to $\theta_{n,0}$ in Θ_n, e.g., $\theta_n := \mathrm{argmin}_{\theta \in \Theta_n} d_n(\theta, \theta_{n,0})$.

Typically, small sieves Θ_n lead to a small modulus, hence fast δ_n in (53). On the other hand, the discrepancy $M_n(\theta_{n,0}) - M_n(\theta_n)$ of a small sieve will be large. Thus, the two terms in the right side of (53) may be loosely understood as a "variance" and a "squared bias" term, which must be balanced to obtain a good rate of convergence. We note that in many problems an un-sieved M-estimator actually performs well, so the trade-off should not be understood too literally: it may work well to reduce the "bias" to zero.

6.1 Least squares regression on sieves

Suppose that we have data

$$Y_i = \theta_0(z_i) + \epsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{54}$$

where $Y_i \in \mathbb{R}$ is the observed response variable, $z_i \in \mathcal{Z}$ is a covariate, and ϵ_i is the unobserved error. The errors are assumed to be independent random variables with expectation $\mathbb{E}\epsilon_i = 0$ and variance $\mathrm{Var}(\epsilon_i) \le \sigma_0^2 < \infty$, for $i = 1, \ldots, n$. The covariates $z_1, \ldots, z_n$ are fixed, i.e., we consider the case of fixed design. The function $\theta_0 : \mathcal{Z} \to \mathbb{R}$ is unknown, but we assume that $\theta_0 \in \Theta$, where Θ is a given class of regression functions.

The unknown regression function can be estimated by the *sieved-least squares estimator* (LSE) $\hat{\theta}_n$, which is defined (not necessarily uniquely) by

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta(z_i))^2,$$

where Θ_n is a set of regression functions $\theta : \mathcal{Z} \to \mathbb{R}$. Inserting the expression for Y_i and calculating the square, we see that $\hat{\theta}_n$ maximizes

$$\mathbb{M}_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} (\theta - \theta_0)(z_i)\epsilon_i - \mathbb{P}_n(\theta - \theta_0)^2,$$

where $\mathbb{P}_n$ is the empirical measure on the design points $z_1, \ldots, z_n$. This criterion function is not observable but is of simpler character than the sum of squares. Note that the second term is assumed non-random, the randomness solely residing in the error terms.

Under the assumption that the error variables have mean zero, the mean of $\mathbb{M}_n(\theta)$ is $M_n(\theta) = -\mathbb{P}_n(\theta - \theta_0)^2$ and can be used as a centering function. It satisfies, for every θ,

$$M_n(\theta) - M_n(\theta_0) = -\mathbb{P}_n(\theta - \theta_0)^2.$$

Thus, Theorem 6.1 applies with $d_n(\theta, \theta_0)$ equal to the $L_2(\mathbb{P}_n)$-distance on the set of regression functions. The modulus of continuity condition takes the form

$$\phi_n(\delta) \ge \mathbb{E} \sup_{\mathbb{P}_n(\theta - \theta_0)^2 \le \delta^2, \theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\theta - \theta_0)(z_i)\epsilon_i \right|. \tag{55}$$

Theorem 6.2. *If $Y_1, \ldots, Y_n$ are independent random variables satisfying (16) for fixed design points $z_1, \ldots, z_n$ and errors $\epsilon_1, \ldots, \epsilon_n$ with mean 0, then the minimizer $\hat{\theta}_n$ over Θ_n of the least squares criterion satisfies*

$$\|\hat{\theta}_n - \theta_0\|_{\mathbb{P}_n, 2} = O_{\mathbb{P}}(\delta_n)$$

for δ_n satisfying $\delta_n \ge \|\theta_0 - \Theta_n\|_{\mathbb{P}_n, 2}$ and $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ for ϕ_n in (55) such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $0 < \alpha < 2$.

Since the design points are non-random, the modulus (55) involves relatively simple multiplier processes, to which the abstract maximal inequalities may apply directly. In particular, if the error variables are sub-Gaussian, then the stochastic process $\{n^{-1/2}\sum_{i=1}^n(\theta-\theta_0)(z_i)\epsilon_i : \theta \in \Theta_n\}$ is sub-Gaussian with respect to the $L_2(\mathbb{P}_n)$- semimetric on the set of regression functions. Thus, using (41), we may choose

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\epsilon, \Theta_n \cap \{\theta : \mathbb{P}_n(\theta-\theta_0)^2 \le \delta^2\}, L_2(\mathbb{P}_n))}\, d\epsilon.$$

Example 6.3 (Bounded isotonic regression)**.** *Let $\Theta_n = \Theta = \{f : [0,1] \to [0,1] : f \text{ is nondecreasing}\}$. By Theorem 2.7.5 of [van der Vaart and Wellner, 1996] we see that*

$$\log N(\epsilon, \Theta, L_2(\mathbb{P}_n)) \le K\epsilon^{-1},$$

where $K > 0$ is a universal constant. Thus, we can take $\phi_n(\delta) = \sqrt{K}\int_0^\delta \epsilon^{-1/2} d\epsilon = 2\sqrt{K}\sqrt{\delta}$. Thus we solve $\sqrt{\delta_n} = \delta_n^2\sqrt{n}$ to obtain the rate of convergence of $\delta_n = n^{-1/3}$.

Example 6.4 (Lipschitz regression)**.** *Let $\Theta = \Theta_n := \{f : [0,1] \to [0,1] \mid f \text{ is 1-Lipschitz}\}$. By Lemma 2.8, we see that $\phi_n(\delta)$ can be taken[48] to be $\sqrt{\delta}$ which yields the rate of $\delta_n = n^{-1/3}$.*

Example 6.5 (Hölder smooth functions)**.** *For $\alpha > 0$, we consider the class of all functions on a bounded set $\mathcal{X} \subset \mathbb{R}^d$ that possess uniformly bounded partial derivatives up to $\lfloor \alpha \rfloor$ and whose highest partial derivates are 'Lipschitz' (actually Hölder) of order $\alpha - \lfloor \alpha \rfloor$[49].*

Let $\mathcal{X} = [0,1]^d$ and let $\Theta_n = C_1^\alpha([0,1]^d)$. Then, $\log N(\epsilon, \Theta, L_2(\mathbb{P}_n)) \le \log N(\epsilon, \Theta, \|\cdot\|_\infty) \lesssim \epsilon^{-d/\alpha}$. Thus, for $\alpha > d/2$ this leads to $\phi_n(\delta) \gg \delta^{1-d/(2\alpha)}$ and hence, $\phi_n(\delta) \le \delta_n^2\sqrt{n}$

[48] Note that a ϵ-cover in the $\|\cdot\|_\infty$-norm (as in Lemma 2.8) also yields a a cover in the $L_2(\mathbb{P}_n)$-seminorm.

[49] i.e., for any vector $\mathbf{k} = (k_1, \ldots, k_d)$ of d integers the differential operator

$$D^{\mathbf{k}} = \frac{\partial^{k.}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k. = \sum_{i=1}^d k_i$. Then for a function $f : \mathcal{X} \to \mathbb{R}$, let

$$\|f\|_\alpha := \max_{k. \le \lfloor\alpha\rfloor} \sup_x \left|D^k f(x)\right| + \max_{k. = \lfloor\alpha\rfloor} \sup_{x,y} \frac{\left|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(y)\right|}{\|x-y\|^{\alpha - \lfloor\alpha\rfloor}},$$

where the supremum is taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_\alpha \le M$. The following lemma, proved in [van der Vaart and Wellner, 1996, Chapter 7], bounds the entropy number of the class $C_M^\alpha(\mathcal{X})$.

Lemma 6.6. *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Then there exists a constant K, depending only on α and d, and a constant K', depending only on α, diam($\mathcal{X}$) and d, such that*

$$\begin{aligned} \log N(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) &\le K\lambda(\mathcal{X}^1)\epsilon^{-d/\alpha}, \\ \log N_{[\,]}(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) &\le K'\epsilon^{-d/\alpha}, \end{aligned}$$

for every $\epsilon > 0$, $r \ge 1$, where $\lambda(\mathcal{X}^1)$ is the Lebesgue measure of the set $\{x : \|x - \mathcal{X}\| \le 1\}$ and Q is any probability measure on $\mathbb{R}^d$. Note that $\|\cdot\|_\infty$ denotes the supremum norm.

can be solved to obtain the rate of convergence $\delta_n \gtrsim n^{-\alpha/(2\alpha+d)}$. The rate relative to the empirical L_2-norm is bounded above by

$$n^{-\alpha/(2\alpha+d)} + \|\theta_0 - \Theta_n\|_{\mathbb{P}_{n,2}}.$$

For $\theta_0 \in C_1^\alpha([0,1]^d)$ the second term vanishes; the first is known to be the minimax rate over this set.

Exercise (HW2) (Convex regression): Suppose that $\theta_0 : C \to \mathbb{R}$ is known to be a convex function over its domain C, some convex and open subset of $\mathbb{R}^d$. In this case, it is natural to consider the LSE with a convexity constraint — namely

$$\hat{\theta}_n \in \underset{f:C\to\mathbb{R} \text{ “convex”}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (Y_i - f(z_i))^2. \tag{56}$$

As stated, this optimization problem is infinite-dimensional in nature. Fortunately, by exploiting the structure of convex functions, it can be converted to an equivalent finite-dimensional problem[50]. Show that the above LSE can be computed by solving the optimization problem:

$$\min_{u_1,\ldots,u_n\in\mathbb{R};\xi_1,\ldots,\xi_n\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (Y_i - u_i)^2 \qquad \text{s.t.} \quad u_i + \xi_i^\top(z_j - z_i) \le u_j \quad \forall i \neq j.$$

Note that this is a convex program in $N = n(d+1)$ variables, with a quadratic cost function and a total of $n(n-1)$ linear constraints. Give the form of a LSE $\hat{\theta}_n$.

Suppose now that $C = [0,1]^d$, and instead of minimizing (56) over the class of all convex functions, we minimize over the class of all L-Lipschitz convex functions. Find the rate of convergence of the LSE (over all L-Lipschitz convex functions).

6.2 Least squares regression: a finite sample inequality

In the *standard nonparametric regression model*, we assume the noise variables in (54) are drawn in an i.i.d. manner from the $N(0,\sigma^2)$ distribution, where $\sigma > 0$ is the unknown standard deviation parameter. In this case, we can write $\epsilon_i = \sigma w_i$, where $w_i \sim N(0,1)$ are i.i.d. We change our notation slightly and assume that $f^* : \mathcal{Z} \to \mathbb{R}$ is the unknown regression function (i.e., $f^* \equiv \theta_0$ in (54)).

[50] Any convex function f is *subdifferentiable* at each point in the (relative) interior of its domain C. More precisely, at any interior point $z \in C$, there exists at least one vector $\xi \in \mathbb{R}^d$ such that

$$f(z) + \xi^\top(x - z) \le f(x), \qquad \text{for all } x \in C.$$

Any such vector is known as a *subgradient*, and each point $z \in C$ can be associated with the set $\partial f(z)$ of its subgradients, which is known as the subdifferential of f at z. When f is actually differentiable at z, then the above inequality holds if and only if $\xi = \nabla f(z)$, so that we have $\partial f(z) = \{\nabla f(z)\}$. See standard references in convex analysis for more on this.

Our main result in this section yields a finite sample inequality for the $L_2(\mathbb{P}_n)$-loss of the constrained LSE

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f(z_i)\}^2;$$

i.e., we study the error $\|\hat{f}_n - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \{\hat{f}_n(z_i) - f^*(z_i)\}^2$. This error is expressed in terms of a *localized form of Gaussian complexity*: it measures the complexity of the function class $\mathcal{F}$, locally in a neighborhood around the true regression function f^*. More precisely, we define the set:

$$\mathcal{F}^* := \mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\} \tag{57}$$

corresponding to an f^*-shifted version of the original function class $\mathcal{F}$. For a given radius $\delta > 0$, the *local Gaussian complexity* around f^* at scale δ is given by

$$G_n(\delta; \mathcal{F}^*) := \mathbb{E}_w \left[\sup_{g \in \mathcal{F}^* : \|g\|_n \le \delta} \left| \frac{1}{n} \sum_{i=1}^{n} w_i g(z_i) \right| \right]$$

where the expectation is w.r.t. the variables $\{w_i\}_{i=1}^n$ which are i.i.d. $N(0,1)$.

A function class $\mathcal{H}$ is *star-shaped* if for any $h \in \mathcal{H}$ and $\alpha \in [0,1]$, the rescaled function αh also belongs to $\mathcal{H}$. Recall the basic inequality for nonparametric least squares:

$$\frac{1}{2}\|\hat{f}_n - f^*\|_n^2 \le \frac{\sigma}{n} \sum_{i=1}^{n} w_i \{f(z_i) - f^*(z_i)\}. \tag{58}$$

A central object in our analysis is the set of $\delta > 0$ that satisfy the *critical inequality*

$$G_n(\delta; \mathcal{F}^*) \le \frac{\delta^2}{2\sigma}. \tag{59}$$

It can be shown that the star-shaped condition ensures existence of the critical radius[51].

[51]Let $\mathcal{H}$ be a star-shaped class of functions.

Lemma 6.7. *For any star-shaped function class $\mathcal{H}$, the function $\delta \mapsto G_n(\delta, \mathcal{H})/\delta$ is nonincreasing on the interval $(0, \infty)$. Consequently, for any constant $c > 0$, the inequality $G_n(\delta, \mathcal{H}) \le c\delta^2$ has a smallest positive solution.*

Proof. For a pair $0 < \delta \le t$, it suffices to show that $\frac{\delta}{t} G_n(t; \mathcal{H}) \le G_n(\delta; \mathcal{H})$. Given any function $h \in \mathcal{H}$ with $\|h\|_n \le t$, we may define the rescaled function $\tilde{h} = \frac{\delta}{t} h$. By construction, we have $\|\tilde{h}\|_n \le \delta$; moreover, since $\delta \le t$, the star-shaped assumption on $\mathcal{H}$ guarantees that $\tilde{h} \in \mathcal{H}$. Thus, write

$$\frac{1}{n} \left| \frac{\delta}{t} \sum_{i=1}^{n} w_i h(z_i) \right| = \frac{1}{n} \left| \sum_{i=1}^{n} w_i \tilde{h}(z_i) \right| \le \sup_{g \in \mathcal{H} : \|g\|_n \le \delta} \frac{1}{n} \left| \sum_{i=1}^{n} w_i g(z_i) \right|.$$

Taking the supremum over the set $\mathcal{H} \cap \{\|h\|_n \le t\}$ on the left-hand side followed by expectations yields $\frac{\delta}{t} G_n(t; \mathcal{H}) \le G_n(\delta; \mathcal{H})$, which completes the proof of the first part. As $G_n(\delta; \mathcal{H})/\delta$ is nonincreasing and $c\delta$ is nondecreasing (in δ) on $(0, \infty)$, the inequality $G_n(\delta, \mathcal{H}) \le c\delta^2$ has a smallest positive solution. $\square$

Theorem 6.8. *Suppose that the shifted function class $\mathcal{F}^*$ is star-shaped, and let δ_n be any positive solution to the critical inequality (59). Then for any $t \geq \delta_n$, the LSE $\hat{f}_n$ satisfies the bound*

$$\mathbb{P}\left(\|\hat{f}_n - f^*\|_n^2 \geq 16t\delta_n\right) \leq e^{-\frac{nt\delta_n}{2\sigma^2}}.$$

Exercise (HW2): By integrating this tail bound, show that the mean-squared error in the $L_2(\mathbb{P}_n)$-semi-norm is upper bounded as

$$\mathbb{E}\left[\|\hat{f}_n - f^*\|_n^2\right] \leq c\left\{\delta_n^2 + \frac{\sigma^2}{n}\right\}$$

for some universal constant c.

Proof. Recall the basic inequality (58). In terms of the shorthand notation $\hat{\Delta} := \hat{f}_n - f^*$, it can be written as $\frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i\hat{\Delta}(z_i)$. By definition, the error function $\hat{\Delta} = \hat{f}_n - f^*$ belongs to the shifted function class $\mathcal{F}^*$. We will need the following lemma.

Lemma 6.9. *Let $\mathcal{H}$ be an arbitrary star-shaped function class, and let $\delta_n > 0$ satisfy the inequality $G_n(\delta; \mathcal{H}) \leq \delta^2/(2\sigma)$. For a given scalar $u \geq \delta_n$, define the event*

$$\mathcal{A}(u) := \left\{\exists\, g \in \{h \in \mathcal{H} : \|h\|_n \geq u\} : \left|\frac{\sigma}{n}\sum_{i=1}^n w_i g(z_i)\right| \geq 2\|g\|_n u\right\}. \tag{60}$$

Then, for all $u \geq \delta_n$, we have

$$\mathbb{P}(\mathcal{A}(u)) \leq e^{-\frac{nu^2}{2\sigma^2}}.$$

We will prove the main theorem using the lemma for the time being; we take $\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$, so that we can write $\mathbb{P}(\mathcal{A}^c(\sqrt{t\delta_n})) \geq 1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}}$. Note that

$$\begin{aligned}
\mathbb{P}(\|\hat{\Delta}\|_n^2 \leq 16t\delta_n) &= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 \leq 16t\delta_n, \|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(\|\hat{\Delta}\|_n^2 \leq 16t\delta_n, \|\hat{\Delta}\|_n^2 \geq t\delta_n\right) \\
&= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2 \leq 16t\delta_n\right) \\
&\geq \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2 \leq 16t\delta_n, \mathcal{A}^c(\sqrt{t\delta_n})\right) \\
&= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2, \mathcal{A}^c(\sqrt{t\delta_n})\right) \qquad (61) \\
&\geq \mathbb{P}\left(\mathcal{A}^c(\sqrt{t\delta_n})\right) \geq 1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}},
\end{aligned}$$

where the only nontrivial step is (61), which we explain next. Note that if $\|\hat{\Delta}\|_n^2 \geq t\delta_n$ and $\mathcal{A}^c(\sqrt{t\delta_n})$ holds, then

$$\left|\frac{1}{n}\sum_{i=1}^n w_i\hat{\Delta}(z_i)\right| \leq 2\|\hat{\Delta}\|_n\sqrt{t\delta_n}.$$

Consequently, the basic inequality (58) implies that $\|\hat{\Delta}\|_n^2 \leq 4\|\hat{\Delta}\|_n\sqrt{t\delta_n}$, or equivalently, $\|\hat{\Delta}\|_n^2 \leq 16t\delta_n$. Thus, (61) holds, thereby completing the proof.

Proof of Lemma 6.9: Our first step is to reduce the problem to controlling a supremum over a subset of functions satisfying the upper bound $\|\tilde{g}\|_n \leq u$. Suppose that there exists some $g \in \mathcal{H}$ with $\|g\|_n \geq u$ such that

$$\Big|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(z_i)\Big| \geq 2\|g\|_n u. \tag{62}$$

Defining the function $\tilde{g} := \frac{u}{\|g\|_n} g$, we observe that $\|\tilde{g}\|_n = u$. Since $g \in \mathcal{H}$ and $\frac{u}{\|g\|_n} \in (0,1]$, the star-shaped assumption on $\mathcal{H}$ implies that $\tilde{g} \in \mathcal{H}$. Consequently, we have shown that if there exists a function g satisfying inequality (62), which occurs whenever the event $\mathcal{A}(u)$ is true, then there exists a function $\tilde{g} \in \mathcal{H}$ with $\|\tilde{g}\|_n = u$ such that

$$\Big|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)\Big| = \frac{u}{\|g\|_n}\Big|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(z_i)\Big| \geq 2u^2.$$

We thus conclude that

$$\mathbb{P}(\mathcal{A}(u)) \leq \mathbb{P}\left(Z_n(u) \geq 2u^2\right), \qquad \text{where} \quad Z_n(u) := \sup_{\tilde{g}\in\mathcal{H}: \|\tilde{g}\|_n \leq u} \Big|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)\Big|.$$

Since the noise variables $w_i \sim N(0,1)$ are i.i.d., the variable $\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)$ is zero-mean and Gaussian for each fixed $\tilde{g}$. Therefore, the variable $Z_n(u)$ corresponds to the supremum of a Gaussian process. If we view this supremum as a function of the standard Gaussian vector $(w_1, \ldots, w_n)$, then it can be verified that the associated Lipschitz constant[52] is at most $\sigma u/\sqrt{n}$. Consequently, by the *concentration of Lipschitz functions of Gaussian variables*[53],

[52] The following lemma illustrates the Lipschitz nature of Gaussian complexity.

Lemma 6.10. *Let $\{W_k\}_{k=1}^n$ be an i.i.d. sequence of $N(0,1)$ variables. Given a collection of vectors $A \subset \mathbb{R}^n$, define the random variable $Z := \sup_{a\in A} |\sum_{k=1}^n a_k W_k|$. Viewing Z as a function $(w_1, \ldots, w_n) \mapsto f(w_1, \ldots, w_n)$, we can verify that f is Lipschitz (with respect to Euclidean norm) with parameter $\sup_{a\in A\cup(-A)} \|a\|_2$.*

To see this, let $w = (w_1, \ldots, w_n), w' = (w_1', \ldots, w_n') \in \mathbb{R}^n$. Suppose that there exists $a^* = (a_1^*, \ldots, a_n^*)$ such that $f(w) = \sup_{a\in A} |\sum_{k=1}^n a_k w_k| = \sum_{k=1}^n a_k^* w_k$ (or $\sum_{k=1}^n (-a_k^*) w_k$, which case can also be handled similarly). Then,

$$f(w) - f(w') \leq \sum_{k=1}^n a_k^* w_k - \sum_{k=1}^n a_k^* w_k' \leq \|a^*\|_2 \|w - w'\|_2 \leq \sup_{a\in A\cup(-A)} \|a\|_2 \|w - w'\|_2.$$

The same argument holds with the roles of w and w' switched which leads to the desired result:

$$|f(w) - f(w')| \leq \sup_{a\in A\cup(-A)} \|a\|_2 \|w - w'\|_2.$$

[53] **Classical result on the concentration properties of Lipschitz functions of Gaussian variables**: Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2, \qquad \text{for all } \ x, y \in \mathbb{R}^n.$$

The following result guarantees that any such function is sub-Gaussian with parameter at most L.

we obtain the tail bound

$$\mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + s\right) \leq e^{-\frac{ns^2}{2u^2\sigma^2}},$$

valid for any $s > 0$. Setting, $s = u^2$ yields,

$$\mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2\right) \leq e^{-\frac{nu^2}{2\sigma^2}}. \tag{63}$$

Finally, by definition of $Z_n(u)$ and $G_n(u; \mathcal{H})$, we have $\mathbb{E}[Z_n(u)] = \sigma G_n(u; \mathcal{H})$. By Lemma 6.7, the function $v \mapsto G_n(v; \mathcal{H})/v$ is nonincreasing, and since $u \geq \delta_n$ by assumption, we have

$$\sigma \frac{G_n(u; \mathcal{H})}{u} \leq \sigma \frac{G_n(\delta_n; \mathcal{H})}{\delta_n} \leq \frac{\delta_n}{2} \leq \delta_n,$$

where the 2nd inequality used the critical condition (59). Putting together the pieces, we have shown that $\mathbb{E}[Z_n(u)] \leq u\delta_n$. Combined with the tail bound (63), we obtain

$$\mathbb{P}(Z_n(u) \geq 2u^2) \leq \mathbb{P}(Z_n(u) \geq u\delta_n + u^2) \leq \mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2\right) \leq e^{-\frac{nu^2}{2\sigma^2}},$$

where we have used the fact that $u^2 \geq u\delta_n$. $\square$

Exercise (HW2): Suppose that $\mathcal{F}^*$ is star-shaped. Show that for any $\delta \in (0, \sigma]$ such that

$$\frac{16}{\sqrt{n}} \int_{\delta^2/(4\sigma)} \sqrt{\log N(t, \mathcal{F}^* \cap \{h : \|h\|_n \leq \delta\}, \|\cdot\|_n)}dt \leq \frac{\delta^2}{4\sigma} \tag{64}$$

satisfies the critical inequality (59) and hence the conclusion of of Theorem 6.8 holds.

Exercise (HW2) [Linear regression]: Consider the standard linear regression model $Y_i = \langle \theta^*, z_i \rangle + w_i$, where $\theta^* \in \mathbb{R}^d$, and fixed x_i are d-dimensional covariates. Although this example can be studied using direct linear algebraic arguments, we will use our general theory in analysis this model. The usual LSE corresponds to optimizing over the class of all linear functions

$$\mathcal{F}_{\text{lin}} := \{f_\theta = \langle \theta, \cdot \rangle : \theta \in \mathbb{R}^d\}. \tag{65}$$

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the design matrix with $z_i \in \mathbb{R}^d$ as its i-th row. Let $\hat{\theta}$ be the LSE. Show that

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \lesssim \sigma^2 \frac{\text{rank}(\mathbf{X})}{n}$$

Theorem 6.11. *Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence*

$$\mathbb{P}\big(|f(X) - \mathbb{E}[f(X)]| \geq t\big) \leq 2e^{-\frac{t^2}{2L^2}} \qquad \text{for all } t \geq 0.$$

Note that this result is truly remarkable: it guarantees that any L-Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2. See Section 13.4 for more details about this result and a proof.

with high probability.

Hint: First note that the shifted class $\mathcal{F}^*_{\text{lin}} = \mathcal{F}_{\text{lin}}$ for any choice of $f_{\theta^*} \in \mathcal{F}_{\text{lin}}$. Moreover, $\mathcal{F}^*_{\text{lin}}$ is convex and hence star-shaped around any point. To use (64) to find δ_n so that Theorem 6.8 applies in this setting, we have to find $N(t, \mathcal{F}^*_{\text{lin}} \cap \{h : \|h\|_n \le \delta\}, \|\cdot\|_n)$. Show that the required covering number can be bounded by $(1 + \frac{2\delta}{t})^r$ where $r := \text{rank}(\mathbf{X})$.

6.3 Oracle inequalities

In our analysis thus far, we have assumed that the true regression function f^* belongs to the function class $\mathcal{F}$ over which the constrained LSE is defined. In practice, this assumption might be violated. In such settings, we expect the performance of the LSE to involve both the *estimation error* that arises in Theorem 6.8, and some additional form of *approximation error*, arising from the fact that $f^* \notin \mathcal{F}$. A natural way in which to measure approximation error is in terms of the best approximation to f^* using functions from $\mathcal{F}$ — the error in this best approximation is given by $\inf_{f\in\mathcal{F}} \|f - f^*\|_n^2$. Note that this error can only be achieved by an "oracle" that has direct access to the samples $\{f^*(x_i)\}_{i=1}^n$. For this reason, results that involve this form of approximation error are referred to as *oracle inequalities*. With this setup, we have the following generalization of Theorem 6.8. We define

$$\partial\mathcal{F} := \{f - g : f, g \in \mathcal{F}\}.$$

Theorem 6.12. *Assume that $\partial\mathcal{F}$ is star-shaped. Let $\delta_n > 0$ be any solution to*

$$G_n(\delta; \partial\mathcal{F}) \le \frac{\delta^2}{2\sigma}. \tag{66}$$

Then for any $t \ge \delta_n$, the LSE $\hat{f}$ satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \le 2 \inf_{f\in\mathcal{F}} \|f - f^*\|_n^2 + 36t\delta_n \tag{67}$$

with probability greater than $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$.

Proof. Recall the definition of $\mathcal{A}(u)$ in (60). We apply Lemma 6.9 with $u = \sqrt{t\delta_n}$ and $\mathcal{H} = \partial\mathcal{F}$ to conclude that $\mathbb{P}\left(\mathcal{A}^c(\sqrt{t\delta_n})\right) \ge 1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$. We will assume below that the event $\mathcal{A}^c(\sqrt{t\delta_n})$ holds.

Given an arbitrary $\tilde{f} \in \mathcal{F}$, since $\tilde{f}$ is feasible and $\hat{f}$ is optimal, we have

$$\frac{1}{2n}\sum_{i=1}^n \{Y_i - \hat{f}(z_i)\}^2 \le \frac{1}{2n}\sum_{i=1}^n \{Y_i - \tilde{f}(z_i)\}^2.$$

Using the relation $Y_i = f^*(z_i) + \sigma w_i$, some algebra yields

$$\frac{1}{2}\|\hat{\Delta}\|_n^2 \le \frac{1}{2}\|\tilde{f} - f^*\|_n^2 + \Big|\frac{\sigma}{n}\sum_{i=1}^n w_i \tilde{\Delta}(z_i)\Big|, \tag{68}$$

where $\hat{\Delta} := \hat{f} - f^*$ and $\tilde{\Delta} := \hat{f} - \tilde{f}$. It remains to analyze the term on the right-hand side involving $\tilde{\Delta}$. We break our analysis into two cases.

Case 1: First suppose that $\|\tilde{\Delta}\|_n \le \sqrt{t\delta_n}$. Then,

$$\begin{aligned}\|\hat{\Delta}\|_n^2 = \|\hat{f} - f^*\|_n^2 &= \|(\tilde{f} - f^*) + \tilde{\Delta}\|_n^2 \\ &\le \left\{\|\tilde{f} - f^*\|_n + \sqrt{t\delta_n}\right\}^2 \\ &\le 2\|\tilde{f} - f^*\|_n^2 + 2t\delta_n \quad (\text{taking } \ \beta = 1)\end{aligned}$$

where in the first inequality above we have used the triangle inequality, and the second inequality follows from the fact that $(a+b)^2 \le 2(a^2+b^2)$ (for $a, b \in \mathbb{R}$).

Case 2: Suppose now that $\|\tilde{\Delta}\|_n > \sqrt{t\delta_n}$. Note that $\tilde{\Delta} \in \partial\mathcal{F}$ and as the event $\mathcal{A}^c(\sqrt{t\delta_n})$ holds, we get

$$\left|\frac{\sigma}{n}\sum_{i=1}^n w_i \tilde{\Delta}(z_i)\right| \le 2\sqrt{t\delta_n}\|\tilde{\Delta}\|_n.$$

Combining with the basic inequality (68), we find that, with probability at least $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$, the squared error is bounded as

$$\begin{aligned}\|\hat{\Delta}\|_n^2 &= \|\tilde{f} - f^*\|_n^2 + 4\sqrt{t\delta_n}\|\tilde{\Delta}\|_n \\ &\le \|\tilde{f} - f^*\|_n^2 + 4\sqrt{t\delta_n}\left\{\|\hat{\Delta}\|_n + \|\tilde{f} - f^*\|_n\right\} \\ &\le \|\tilde{f} - f^*\|_n^2 + 2\left[\frac{t\delta_n}{\beta} + \beta\|\hat{\Delta}\|_n^2\right] + 2\left[\frac{t\delta_n}{\beta} + \beta\|\tilde{f} - f^*\|_n^2\right] \\ \Rightarrow \quad (1-2\beta)\|\hat{\Delta}\|_n^2 &\le (1+2\beta)\|\tilde{f} - f^*\|_n^2 + 4\frac{t\delta_n}{\beta}\end{aligned}$$

where the second step follows from the triangle inequality and the next step follows from multiple usage of the fact that $2ab \le \beta a^2 + b^2/\beta$ (for $a, b \in \mathbb{R}$ and $\beta > 0$). Taking $\beta = 1/6$, we have $\frac{(1+2\beta)}{(1-2\beta)} = 2$, and thus we get

$$\|\hat{\Delta}\|_n^2 \le 2\|\tilde{f} - f^*\|_n^2 + 36t\delta_n.$$

Combining the pieces we get that, under the event $\mathcal{A}^c(\sqrt{t\delta_n})$, the above inequality holds for any $\tilde{f} \in \mathcal{F}$. Thus, (67) holds. $\square$

Remark 6.1. *We can, in fact, have a slightly more general form of* (67) *where the 'oracle' approximation term* $2\|\tilde{f} - f^*\|_n^2$ *can be replaced by* $\frac{1+\gamma}{1-\gamma}\|\tilde{f} - f^*\|_n^2$ *for any* $\gamma \in (0,1)$ *(with appropriate adjustments to the 'estimation' error term* $36t\delta_n$*).*

Note that the guarantee (67) is actually a family of bounds, one for each $f \in \mathcal{F}$. When $f^* \in \mathcal{F}$, then we can set $f = f^*$, so that the bound (67) reduces to asserting that $\|\hat{f} - f^*\|_n^2 \lesssim t\delta_n$ with high probability, where δ_n satisfies the critical inequality (66). Thus, up to constant factors, we recover Theorem 6.8 as a special case of Theorem 6.12. By integrating the tail bound, we are guaranteed that

$$\mathbb{E}\left[\|\hat{f} - f^*\|_n^2\right] \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2 + \frac{\sigma^2}{n}. \tag{69}$$

The bound (69) guarantees that the LSE $\hat{f}$ has prediction error that is at most a constant multiple of the oracle error, plus a term proportional to δ_n^2. The term $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$ can be viewed a form of approximation error that decreases as the function class $\mathcal{F}$ grows, whereas the term δ_n^2 is the estimation error that increases as $\mathcal{F}$ becomes more complex.

6.3.1 Best sparse linear regression

Consider the standard linear model $Y_i = f_{\theta^*}(z_i) + \sigma w_i$, where $f_\theta(z) := \langle \theta, z \rangle$ is an unknown linear regression function, and $w_i \overset{iid}{\sim} N(0,1)$ is an i.i.d. noise sequence. Here $\theta^* \in \mathbb{R}^d$ is the unknown parameter. For some sparsity index $s \in \{1, 2, \ldots, d\}$, consider the class of all linear regression functions based on s-sparse vectors — namely, the class

$$\mathcal{F}_{\text{spar}}(s) := \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\|_0 \le s\},$$

where $\|\theta\|_0 := \sum_{j=1}^d I(\theta_j \neq 0)$ counts the number of non-zero coefficients in the vector $\theta \in \mathbb{R}^d$. Disregarding computational considerations, a natural estimator of θ^* is given by

$$\hat{\theta} \equiv f_{\hat{\theta}} \in \arg \min_{f_\theta \in \mathcal{F}_{\text{spar}(s)}} \sum_{i=1}^n \{Y_i - f_\theta(z_i)\rangle\}^2, \tag{70}$$

corresponding to performing least squares over the set of all regression vectors with at most s non-zero coefficients. As a corollary of Theorem 6.12, we claim that the $L_2(\mathbb{P}_n)$-error of this estimator is upper bounded as

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 \lesssim \inf_{\theta \in \mathcal{F}_{\text{spar}(s)}} \|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 + \sigma^2 \frac{s \log(\frac{ed}{s})}{n}, \tag{71}$$

with high probability; here $\delta_n^2 = \sigma^2 \frac{s \log(\frac{ed}{s})}{n}$. Consequently, up to constant factors, its error is as good as the best s-sparse predictor plus the 'estimation' error term δ_n^2. Note that this 'estimation' error term grows linearly with the sparsity s, but only logarithmically in the dimension d, so that it can be very small even when the dimension is exponentially larger than the sample size n. In essence, this result guarantees that we pay a relatively small price for not knowing in advance the best s-sized subset of coefficients to use.

In order to derive this result as a consequence of Theorem 6.12, we need to compute the local Gaussian complexity $G_n(\delta; \partial\mathcal{F}_{\text{spar}}(s))$. Making note of the inclusion $\partial\mathcal{F}_{\text{spar}}(s) \subset \mathcal{F}_{\text{spar}}(2s)$, we have $G_n(\delta; \partial\mathcal{F}_{\text{spar}}(s)) \subset G_n(\delta; \mathcal{F}_{\text{spar}}(2s))$. Now let $S \subset \{1, \ldots, d\}$ be an arbitrary $2s$-sized subset of indices, and let $\mathbf{X}_S \in \mathbb{R}^{n \times 2s}$ denote the submatrix with columns indexed by S. We can then write

$$G_n(\delta; \mathcal{F}_{\text{spar}}(2s)) = \mathbb{E}_w \left[\sup_{g \in \mathcal{F}_{\text{spar}}(2s): \|g\|_n \le \delta} \left| \frac{1}{n} \sum_{i=1}^n w_i g(z_i) \right| \right] = \mathbb{E}_w \left[\max_{|S|=2s} Z_n(S) \right],$$

where

$$Z_n(S) := \sup_{\theta_S \in \mathbb{R}^{2s}: \frac{\|\mathbf{X}_S \theta_S\|_2}{\sqrt{n}} \le \delta} \frac{1}{n} \left| w^\top \mathbf{X}_S \theta_S \right|$$

as, for $g \in \mathcal{F}_{\text{spar}}(2s)$, $g(z) \equiv g_\theta(z) = \langle \theta, z \rangle = \langle \theta_S, z_S \rangle$, if θ has nonzero entries in the subset $S \subset \{1, \ldots, d\}$, and $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \theta, z_i \rangle^2 = \frac{1}{n}\|\mathbf{X}_S \theta_S\|_2^2$ (here $\|\cdot\|_2$ denotes the usual Euclidean norm).

Viewed as a function of the standard Gaussian vector $w \in \mathbb{R}^n$, the variable $Z_n(S)$ is Lipschitz with parameter at most $\delta/\sqrt{n}$ (by Lemma 6.10), from which Theorem 6.11 implies the tail bound

$$\mathbb{P}(Z_n(S) \geq \mathbb{E}[Z_n(S)] + t\delta) \leq e^{\frac{-t^2\delta^2}{(2\delta^2/n)}} = e^{\frac{-nt^2}{2}}, \qquad \text{for all } t > 0. \tag{72}$$

We now upper bound the expectation. Consider the singular value decomposition $\mathbf{X}_S = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times 2s}$ and $\mathbf{V} \in \mathbb{R}^{d\times 2s}$ are matrices of left and right singular vectors, respectively, and $\mathbf{D} \in \mathbb{R}^{2s\times 2s}$ is a diagonal matrix of the singular values. Noting that $\|\mathbf{X}_S\theta_S\|_2 = \|\mathbf{D}\mathbf{V}^\top\theta_S\|_2$, we arrive at the upper bound

$$\mathbb{E}[Z_n(S)] \leq \mathbb{E}\Big[\sup_{\beta\in\mathbb{R}^{2s}:\|\beta\|_2\leq\delta} \frac{1}{\sqrt{n}}\big|\langle \mathbf{U}^\top w, \beta\rangle\big|\Big] \leq \frac{\delta}{\sqrt{n}}\mathbb{E}\left[\|\mathbf{U}^\top w\|_2\right]$$

where we have taken $\beta = \frac{\mathbf{D}\mathbf{V}^\top\theta_S}{\sqrt{n}}$. Since $w \sim N(0, I_n)$ and the matrix $\mathbf{U}$ has orthogonal columns, we have $\mathbf{U}^\top w \sim N(0, I_{2s})$, and therefore $\mathbb{E}\left[\|\mathbf{U}^\top w\|_2\right] \leq \sqrt{2s}$. Combining this upper bound with the earlier tail bound (72), an application of the union bound yields, for all $t > 0$,

$$\mathbb{P}\left[\max_{|S|=2s} Z_n(S) \geq \frac{\delta\sqrt{2s}}{\sqrt{n}} + t\delta\right] \leq \binom{d}{2s} e^{\frac{-nt^2}{2}}.$$

By integrating this tail bound, we find that

$$\frac{G_n(\delta; \mathcal{F}_{\text{spar}}(2s))}{\delta} = \frac{\mathbb{E}_w\left[\max_{|S|=2s} Z_n(S)\right]}{\delta} \lesssim \sqrt{\frac{s}{n}} + \sqrt{\frac{\log\binom{d}{2s}}{n}} \lesssim \sqrt{\frac{\log\frac{ed}{s}}{n}},$$

so that the critical inequality (66) is satisfied for $\delta_n^2 \simeq \sigma^2 \frac{s\log(\frac{ed}{s})}{n}$, as claimed.

6.4 Density estimation via maximum likelihood

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a density p_0 that belongs to a set $\mathcal{P}$ of densities with respect to a measure μ on some measurable space. In this subsection the parameter is the density p_0 itself (and we denoted a generic density by p instead of θ).

The *sieved maximum likelihood estimator* (MLE) $\hat{p}_n$ based on $X_1, \ldots, X_n$ maximizes the log-likelihood $p \mapsto \mathbb{P}_n \log p$ over a sieve $\mathcal{P}_n$, i.e.,

$$\hat{p}_n = \operatorname*{argmax}_{p\in\mathcal{P}_n} \mathbb{P}_n[\log p].$$

Although it is natural to take the objective (criterion) function we optimize (i.e., $\mathbb{M}_n(\cdot)$ in our previous notation) as $\mathbb{P}_n \log p$, for some technical reasons (explained below) we consider a slightly modified function.

Let $p_n \in \mathcal{P}_n$. We will discuss the particular choice of p_n later. By concavity of the logarithm function we have

$$\mathbb{P}_n \log \frac{\hat{p}_n + p_n}{2p_n} \geq \mathbb{P}_n[\frac{1}{2} \log \frac{\hat{p}_n}{p_n} + \frac{1}{2} \log 1] \geq 0 = \mathbb{P}_n \log \frac{p_n + p_n}{2p_n}.$$

Thus, defining the criterion functions $m_{n,p}$ (for $p \in \mathcal{P}_n$) as

$$m_{n,p} := \log \frac{p + p_n}{2p_n},$$

we obtain $\mathbb{P}_n m_{n,\hat{p}_n} \geq \mathbb{P}_n[m_{n,p_n}]$. We shall apply Theorem 6.1 with $\mathbb{M}_n(p) := \mathbb{P}_n[m_{n,p}]$ to obtain the rate of convergence of $\hat{p}_n$. We note that it is not true that $\hat{p}_n$ maximizes the map $p \mapsto \mathbb{M}_n(p)$ over $\mathcal{P}_n$. Inspection of the conditions of Theorem 6.1 shows that this is not required for its application; it suffices that the criterion is bigger at the estimator than at the value θ_n, which is presently taken equal to p_n.

An immediate question that arises next is what discrepancy (or metric) do we use to measure the difference between $\hat{p}_n$ and p_0? A natural metric while comparing densities is the Hellinger distance:

$$h(p, q) = \left[\int (\sqrt{p} - \sqrt{q})^2 d\mu\right]^{1/2},$$

which is what we will use in this subsection.

We will apply Theorem 6.1 with $\theta_{n,0} = \theta_n = p_n$ in our new notation. Thus, we will obtain a rate of convergence for $h(\hat{p}_n, p_n)$, which coupled with a satisfactory choice of p_n will yield a rate for $h(\hat{p}_n, p_0)$. It is then required that the centering function decreases quadratically as p moves away from p_n within the sieve, at least for $h(p, p_n) > \tilde{\delta}_n$ (to be defined later). For $\delta > 0$, let

$$\mathcal{M}_{n,\delta} := \{m_{n,p} - m_{n,p_n} : p \in \mathcal{P}_n, h(p, p_n) \leq \delta\}.$$

We will need to use a maximal inequality to control the fluctuations of the empirical process in the class $\mathcal{M}_{n,\delta}$. The following result will be useful in this regard; it uses bracketing with the "Bernstein norm"[54].

Theorem 6.13. *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{P,B} < \delta$ for every f,*

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|_{P,B}) \left(1 + \frac{J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})}{\delta^2 \sqrt{n}}\right).$$

Using $m_{n,p}$ rather than the more obvious choice $\log p$ is technically more convenient. First it combines smoothly with the Hellinger distance $h(p, q)$. The key is the following pair of inequalities, which relate the "Bernstein norm" of the criterion functions $m_{n,p}$ to the Hellinger distance of the densities p.

[54] The "Bernstein norm" is defined as $\|f\|_{P,B} := \left[2P(e^{|f|} - 1 - |f|)\right]^{1/2}$. This "Bernstein norm" turns out to combine well with minimum contrast estimators (e.g., MLE), where the criterion is a logarithm of another natural function, such as the log-likelihood. Actually, $\|f\|_{P,B}$ is not a true norm, but it can be used in the same way to measure the size of brackets.

Lemma 6.14. *For nonnegative functions p, q, p_n, and p_0 (assumed to be a density) such that $p_0/p_n \leq M$ and $p \leq q$, we have*

$$\|m_{n,p} - m_{n,p_n}\|_{P_0,B} \lesssim \sqrt{M} h(p, p_n), \qquad \|m_{n,p} - m_{n,q}\|_{P_0,B} \lesssim \sqrt{M} h(p, q),$$

where the constant in $\lesssim$ does not depend on anything.

Proof. Note that $e^{|x|} - 1 - |x| \leq 4(e^{x/2} - 1)^2$, for every $x \geq -\log 2$. As $m_{n,p} \geq -\log 2$ and $m_{n,p_n} = 0$,

$$\|m_{n,p} - m_{n,p_n}\|^2_{P_0,B} \lesssim P_0 \Big(e^{m_{n,p}/2} - 1\Big)^2 = P_0 \left(\frac{\sqrt{p + p_n}}{\sqrt{2p_n}} - 1\right)^2.$$

Since $p_0/p_n \leq M$, the right side is bounded by $2Mh^2(p + p_n, 2p_n)$. Combination with the preceding display gives the first inequality. If $p \leq q$, then $m_{n,q} - m_{n,p}$ is nonnegative. By the same inequality for $e^x - 1 - x$ as before,

$$\|m_{n,p} - m_{n,q}\|^2_{P_0,B} \lesssim P_0 \Big(e^{(m_{n,q} - m_{n,p})/2} - 1\Big)^2 = P_0 \left(\frac{\sqrt{q + p_n}}{\sqrt{p + p_n}} - 1\right)^2.$$

This is bounded by $Mh^2(q + p_n, p + p_n) \lesssim Mh^2(p, q)$ as before. □

Since the map $p \mapsto m_{n,p}$ is monotone, the second inequality shows that a bracketing partition of a class of densities p for the Hellinger distance induces a bracketing partition of the class of criterion functions $m_{n,p}$ for the "Bernstein norm" of essentially the same size. Thus, we can use a maximal inequality available for the classes of functions $\mathcal{M}_{n,\delta}$ with the entropy bounded by the Hellinger entropy of the class of densities.

Lemma 6.15. *Let h denote the Hellinger distance on a class of densities $\mathcal{P}$ and set $m_{n,p} := \log[(p + p_n)/(2p_n)]$. If p_n and p_0 are probability densities with $p_0/p_n \leq M$ pointwise, then*

$$P_0[m_{n,p} - m_{n,p_n}] \lesssim -h^2(p, p_n),$$

for every probability density p such that $h(p, p_n) \geq ch(p_n, p_0)$, for some constant $c > 0$. Furthermore, for the class of functions $\mathcal{M}_{n,\delta} := \{m_{n,p} - m_{n,p_n} : p \in \mathcal{P}_n, h(p, p_n) \leq \delta\}$,

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim \sqrt{M} J_{[\,]}(\delta, \mathcal{P}_n, h) \left(1 + \frac{\sqrt{M} J_{[\,]}(\delta, \mathcal{P}_n, h)}{\delta^2 \sqrt{n}}\right).$$

Proof. Since $\log x \leq 2(\sqrt{x} - 1)$ for every $x > 0$,

$$\begin{aligned} P_0 \log \frac{q}{p_n} &\leq 2P_0 \left(\frac{q^{1/2}}{p_n^{1/2}} - 1\right) \\ &= 2P_n \left(\frac{q^{1/2}}{p_n^{1/2}} - 1\right) + 2 \int (q^{1/2} - p_n^{1/2})(p_0^{1/2} - p_n^{1/2}) \frac{p_0^{1/2} + p_n^{1/2}}{p_n^{1/2}} \, d\mu. \end{aligned}$$

The first term in last display equals $-h^2(q,p_n)$. The second term can be bounded by the expression $2h(q,p_n)h(p_0,p_n)(\sqrt{M}+1)$ in view of the assumption on the quotient p_0/p_n and the Cauchy-Schwarz inequality. The sum is bounded by $-h^2(q,p_n)/2$ if $2h(p_0,p_n)(\sqrt{M}+1) \le h(q,p_n)/2$. The first statement of the theorem follows upon combining this with the inequalities[55] [Exercise (HW2)]

$$h(2p, p+q) \le h(p,q) \le (1+\sqrt{2})\, h(2p, p+q).$$

These inequalities are valid for every pair of densities p and q and show that the Hellinger distance between p and q is equivalent to the Hellinger distance between p and $(p+q)/2$.

The maximal inequality is now a consequence of Theorem 6.13. Each of the functions in $\mathcal{M}_{n,\delta}$ has "Bernstein norm" bounded by a multiple of $\sqrt{M}\delta$, while a bracket $[p^{1/2}, q^{1/2}]$ of densities of size δ leads to a bracket $[m_{n,p}, m_{n,q}]$ of "Bernstein norm" of size a multiple of $\sqrt{M}\delta$. □

It follows that the conditions of Theorem 6.1 are satisfied with the Hellinger distance, $\tilde{\delta}_n = h(p_n, p_0)$, and

$$\phi_n(\delta) := J_{[\,]}(\delta, \mathcal{P}_n, h)\left(1 + \frac{J_{[\,]}(\delta, \mathcal{P}_n, h)}{\delta^2\sqrt{n}}\right),$$

where $J_{[\,]}(\delta, \mathcal{P}_n, h)$ is the Hellinger bracketing integral of the sieve $\mathcal{P}_n$. (Usually this function $\phi_n(\cdot)$ has the property that $\phi_n(\delta)/\delta^\alpha$ is decreasing for some $0<\alpha<2$ as required by Theorem 6.1.) The condition $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n$ is equivalent to

$$J_{[\,]}(\delta_n, \mathcal{P}_n, h) \le \sqrt{n}\delta_n^2.$$

For the unsieved MLE the Hellinger integral is independent of n and any δ_n solving the preceding display gives an upper bound on the rate. Under the condition that the true density p_0 can be approximated by a sequence $p_n \in \mathcal{P}_n$ such that p_0/p_n is uniformly bounded, the sieved MLE that maximizes the likelihood over $\mathcal{P}_n$ has at least the rate δ_n satisfying both

$$J_{[\,]}(\delta_n, \mathcal{P}_n, h) \le \sqrt{n}\delta_n^2 \qquad \text{and} \qquad \delta_n \gtrsim h(p_n, p_0).$$

Theorem 6.16. *Given a random sample $X_1, \ldots, X_n$ from a density p_0 let $\hat{p}_n$ maximize the likelihood $p \mapsto \prod_{i=1}^n p(X_i)$ over an arbitrary set of densities $\mathcal{P}_n$. Then*

$$h(\hat{p}_n, p_0) = O_P(\delta_n)$$

for any δ_n satisfying

$$J_{[\,]}(\delta_n, \mathcal{P}_n, h) \le \sqrt{n}\delta_n^2, \qquad \textit{and} \qquad \delta_n \gtrsim h(p_n, p_0)$$

[55] The inequalities follow, because for any nonnegative reals s and t,

$$|\sqrt{2s} - \sqrt{s+t}| \le |\sqrt{s} - \sqrt{t}| \le (1+\sqrt{2})|\sqrt{2s} - \sqrt{s+t}|.$$

The lower inequality follows from the concavity of the root function. The upper inequality is valid with constant $\sqrt{2}$ if $t \ge s$ and with constant $(1+\sqrt{2})$ as stated if $t \le s$.

where p_n can be any sequence with $p_n \in \mathcal{P}_n$ for every n and such that the functions $x \mapsto p_0(x)/p_n(x)$ are uniformly bounded in x and n.

Example 6.17. *Suppose the observations take their values in a compact interval $[0, T]$ in the real line and are sampled from a density that is known to be nonincreasing. Conclude that if $\mathcal{P}$ is the set of all nonincreasing probability densities bounded by a constant C, then*

$$\log N_{[\,]}(\epsilon, \mathcal{P}, h) \leq \log N_{[\,]}(\epsilon, \mathcal{F}, L_2(\lambda)) \lesssim \frac{1}{\epsilon}.$$

where $\mathcal{F}$ of all non-increasing functions $f : [0, T] \to [0, \sqrt{C}]$. The result follows from the observations: (i) $\mathcal{F}$ has bracketing entropy for the $L_2(\lambda)$-norm of the order $1/\epsilon$ for any finite measure λ on $[0, T]$, in particular the Lebesgue measure; (ii) if a density p is non-increasing, then so is its root $\sqrt{p}$; (iii) the Hellinger distance on the densities is the $L_2(\lambda)$-distance on the root densities.

Thus $J_{[\,]}(\delta, \mathcal{P}, h) \lesssim \sqrt{\delta}$, which yields a rate of convergence of at least $\delta_n = n^{-1/3}$ for the MLE. The MLE is called the Grenander estimator.

7 Vapnik-Červonenkis (VC) classes of sets/functions

Consider our canonical setting: $X_1, \ldots, X_n$ are i.i.d. P on some space $\mathcal{X}$. In this section we study classes of functions $\mathcal{F}$ (on $\mathcal{X}$) that satisfy certain *combinatorial restrictions*. These classes at first sight may seem have nothing to do with entropy numbers, but indeed will be shown to imply bounds on the covering numbers of the type

$$\sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^V, \qquad 0 < \epsilon < 1, \text{ some number } V > 0,$$

where $\mathcal{F}$ is the underlying function class with envelope F, and K is a universal constant. Note that this has direct implications on the uniform entropy of such a class (see Definition 4.6) is of the order $\log(1/\epsilon)$ and hence the uniform entropy integral converges, and is of the order $\delta \log(1/\delta)$, as $\delta \downarrow 0$.

Classes of (indicator functions of) this type were first studied by Vapnik and Červonenkis in the 1970s, whence the name VC classes. There are many examples of VC classes, and more examples can be constructed by operations as unions and sums. Furthermore, one can combine VC classes in different sorts of ways (thereby, building larger classes of functions) to ensure that the resulting larger classes also satisfy the uniform entropy condition (though these larger classes may not necessarily be VC).

We first consider VC classes to sets. To motivate this study let us consider a boolean class of functions $\mathcal{F}$[56], i.e., every $f \in \mathcal{F}$ takes values in $\{0, 1\}$. Thus,

$$\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\},$$

where $\mathcal{C}$ is a collection of subsets of $\mathcal{X}$. This naturally leads to the study of $\mathcal{C}$.

Definition 7.1. *Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. Let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be an arbitrary set of n points. Say that $\mathcal{C}$ picks out a certain subset A of $\{x_1, \ldots, x_n\}$ if A can be expressed as $C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$.*

The collection $\mathcal{C}$ is said to shatter $\{x_1, \ldots, x_n\}$ if each of its 2^n subsets can be picked out in this manner (note that an arbitrary set of n points possesses 2^n subsets).

Definition 7.2. *The VC dimension $V(\mathcal{C})$ of the class $\mathcal{C}$ is the largest n such that some set of size n is shattered by $\mathcal{C}$.*

Definition 7.3. *The VC index $\Delta_n(\mathcal{C}; x_1, \ldots, x_n)$ is defined as*

$$\Delta_n(\mathcal{C}; x_1, \ldots, x_n) = \left|\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}\right|,$$

where $|A|$ denotes the cardinality of the set A. Thus,

$$V(\mathcal{C}) := \sup\{n : \max_{x_1,\ldots,x_n \in \mathcal{X}} \Delta_n(\mathcal{C}; x_1, \ldots, x_n) = 2^n\}.$$

[56] Boolean classes $\mathcal{F}$ arise in the problem of classification (where $\mathcal{F}$ can be taken to consist of all functions f of the form $I\{g(X) \neq Y\}$). They are also important for historical reasons: empirical process theory has its origins in the study of the function class $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$.

Figure 1: The left panel illustrates how we can pick out 2 points and 3 points (it's clear that capturing just 1 point and all 4 points are both trivial) thereby showing that there exist 4 points that can be shattered. The right panel illustrates that no set of 5 points can be shattered: the minimum enclosing rectangle that allows us to select all 5 points is defined by only four points — one for each edge. So, it is clear that the fifth point must lie either on an edge or on the inside of the rectangle thereby preventing us from selecting four points without the fifth.

Clearly, the more refined $\mathcal{C}$ is, higher the VC index. The VC dimension is infinite if $\mathcal{C}$ shatters sets of arbitrarily large size. It is immediate from the definition that $V(\mathcal{C}) \leq V$ if and only if no set of size $V + 1$[57] is shattered.

Example 7.4. *Let $\mathcal{X} = \mathbb{R}$ and define the collection of sets $\mathcal{C} := \{(-\infty, c] : c \in \mathbb{R}\}$. Consider any two point set $\{x_1, x_2\} \subset \mathbb{R}$, and assume without loss of generality, that $x_1 < x_2$. It is easy to verify that $\mathcal{C}$ can pick out the null set $\{\}$ and the sets $\{x_1\}$ and $\{x_1, x_2\}$ but cannot pick out $\{x_2\}$. Hence its VC dimension equals 1.*

The collection of all cells $(a, b] \in \mathbb{R}$ shatters every two-point set but cannot pick out the subset consisting of the smallest and largest points of any set of three points. Thus its VC dimension equals 2.

Remark 7.1. *With more effort, it can be seen that VC dimensions of the same type of sets in $\mathbb{R}^d$ are d and $2d$, respectively. For example, let $\mathcal{X} = \mathbb{R}^2$ and define*

$$\mathcal{C} = \{A \subset \mathcal{X} : A = [a, b] \times [c, d], \text{ for some } a, b, c, d \in \mathbb{R}\}.$$

Let us see what happens when $n = 4$. Draw a figure to see this when the points are not co-linear. We can show that there exists 4 points such that all the possible subsets of these four points are picked out by $\mathcal{C}$; see the left panel of Figure 7.1.

Now if we have $n = 5$ points things change a bit; see the right panel of Figure 7.1. If we have five points there is always one that stays "in the middle" of all the others, and thus the complement set cannot be picked out by $\mathcal{C}$. We immediately conclude that the VC dimension of $\mathcal{C}$ is 4.

A collection of measurable sets $\mathcal{C}$ is called a *VC class* if its dimension is finite. The main result of this section is the remarkable fact that the covering numbers of any VC class grow polynomially in $1/\epsilon$ as $\epsilon \to 0$, of order dependent on the dimension of the class.

[57] Some books define the *VC index* of the class $\mathcal{C}$ as the smallest n for which no set of size n is shattered by $\mathcal{C}$ (i.e., $V(\mathcal{C}) + 1$ in our notation).

Example 7.5. *Suppose that $\mathcal{X} = [0,1]$, and let $\mathcal{C}$ be the class of all finite subsets of $\mathcal{X}$. Let P be the uniform (Lebesgue) distribution on $[0,1]$. Clearly $V(\mathcal{C}) = \infty$ and $\mathcal{C}$ is not a VC class. Note that for any possible value of $\mathbb{P}_n$ we have $\mathbb{P}_n(A) = 1$ for $A = \{X_1, \ldots, X_n\}$ while $P(A) = 0$. Therefore $\|\mathbb{P}_n - P\|_{\mathcal{C}} = 1$ for all n, so $\mathcal{C}$ is not a Glivenko-Cantelli class for P.*

Exercise (HW3): Show that the class of all closed and convex sets in $\mathbb{R}^d$ does not have finite VC dimension (Hint: Consider a set of n points on the boundary of the unit ball).

Sauer's lemma[58] (also known as Sauer-Shelah-Vapnik-Červonenkis lemma), one of the fundamental results on VC dimension, states that the number $\Delta_n(\mathcal{C}; x_1, \ldots, x_n)$ of subsets picked out by a VC class $\mathcal{C}$, for $n \geq 1$, satisfies:

$$\max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}; x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})} \binom{n}{j}, \tag{73}$$

where we use the notation $\binom{n}{j} = 0$ if $j > n$. Observe that for $n \leq V(\mathcal{C})$, the right-hand side of the above display equals 2^n, i.e., the growth is exponential. However, it is easy to show[59] that for $n \geq V(\mathcal{C})$,

$$\sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} \leq \left(\frac{ne}{V(\mathcal{C})}\right)^{V(\mathcal{C})}. \tag{74}$$

Consequently, the numbers on the left side grow polynomially (of order at most $O(n^{V(\mathcal{C})})$) rather than an exponential number. Intuitively this means that a finite VC index implies that $\mathcal{C}$ has an apparent simplistic structure.

7.1 VC classes of Boolean functions

The definition of VC dimension can be easily extended to a function class $\mathcal{F}$ in which every function f is binary-valued, taking the values $\{0, 1\}$ (say). In this case, we define, for every

[58]See [van der Vaart and Wellner, 1996, pages 135–136] for a complete proof of the result.

[59]In the following we just give a proof of the right-hand inequality of (74). Note that with $Y \sim$ Binomial$(n, 1/2)$,

$$\begin{aligned}
\sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} &= 2^n \sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} \left(\frac{1}{2}\right)^n = 2^n \mathbb{P}(Y \leq V(\mathcal{C})) \\
&\leq 2^n \mathbb{E}[r^{Y - V(\mathcal{C})}] \qquad \text{for } r \leq 1 \qquad \left[\text{as } \mathbf{1}\{Y - V(\mathcal{C}) \leq 0\} \leq r^{Y - V(\mathcal{C})} \text{ for } r \leq 1\right] \\
&= 2^n r^{-V(\mathcal{C})} \left(\frac{1}{2} + \frac{r}{2}\right)^n = r^{-V(\mathcal{C})}(1+r)^n \qquad \left[\text{as } \mathbb{E}[r^Y] = \sum_{j=0}^{n} r^j \binom{n}{j} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2} + \frac{r}{2}\right)^n\right] \\
&= \left(\frac{n}{V(\mathcal{C})}\right)^{V(\mathcal{C})} \left(1 + \frac{V(\mathcal{C})}{n}\right)^n \qquad \text{by choosing } r = V(\mathcal{C})/n \\
&\leq \left(\frac{n}{V(\mathcal{C})}\right)^{V(\mathcal{C})} e^{V(\mathcal{C})}.
\end{aligned}$$

$x_1, \ldots, x_n \in \mathcal{X}$,

$$\mathcal{F}(x_1, \ldots, x_n) := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}. \tag{75}$$

As functions in $\mathcal{F}$ are Boolean, $\mathcal{F}(x_1, \ldots, x_n)$ is a subset of $\{0,1\}^n$.

Definition 7.6. *Given such a function class $\mathcal{F}$ we say that the set $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{F}$ if*

$$\Delta_n(\mathcal{F}; x_1, \ldots, x_n) := |\mathcal{F}(x_1, \ldots, x_n)| = 2^n.$$

The VC dimension $V(\mathcal{F})$ of $\mathcal{F}$ is defined as the largest integer n for which there is some collection $x_1, \ldots, x_n$ of n points that can be shattered by $\mathcal{F}$.

When $V(\mathcal{F})$ is finite, then $\mathcal{F}$ is said to be a VC class.

Example 7.7. *Let us revisit the Glivenko-Cantelli (GC) theorem (Theorem 3.5) when we have a binary-valued function class $\mathcal{F}$. In particular, suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. A natural question is how does one verify condition (8) in practice? We need an upper bound on $N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))$. Recall that under $L_1(\mathbb{P}_n)$ the distance between f and g is measured by*

$$\|f - g\|_{L_1(\mathbb{P}_n)} := \frac{1}{n}\sum_{i=1}^n |f(X_i) - g(X_i)|.$$

This notion of distance clearly only depends on the values of f and g at the data points $X_1, \ldots, X_n$. Therefore, the covering number of $\mathcal{F}$ in the $L_1(\mathbb{P}_n)$-norm should be bounded from above by the corresponding covering number of $\{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$. It should be obvious that $N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))$ is bounded from above by the cardinality of $\mathcal{F}(X_1, \ldots, X_n)$, i.e.,

$$N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n)) \le |\mathcal{F}(X_1, \ldots, X_n)| \qquad \text{for every } \epsilon > 0.$$

This is in fact a very crude upper bound although it can be quite useful in practice. For example, in the classical GC theorem $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$, and we can see that $|\mathcal{F}(X_1, \ldots, X_n)| \le (n+1)$.

Since $\mathcal{F}(X_1, \ldots, X_n)$ is a subset of $\{0,1\}^n$, its maximum cardinality is 2^n. But if $\Delta_n(\mathcal{F}; X_1, \ldots, X_n)$ is at the most a polynomial in n for every possible realization of $X_1, \ldots, X_n$, then

$$\frac{1}{n}\log \Delta_n(\mathcal{F}; X_1, \ldots, X_n) \to 0 \qquad \text{as } n \to \infty \text{ a.s.} \tag{76}$$

which implies, by Theorem 3.5, that $\mathcal{F}$ is GC. Thus, if $\mathcal{F}$ is a boolean function class such that (76) holds, then $\mathcal{F}$ is P-GC.

Exercise (HW3): Consider the class of all two-sided intervals over the real line, i.e., $\mathcal{F} := \{\mathbf{1}_{(a,b]}(\cdot) : a < b \in \mathbb{R}\}$. Show that $\Delta_n(\mathcal{F}; X_1, \ldots, X_n) \le (n+1)^2$ a.s.

Exercise (HW3): For a scalar $t \in \mathbb{R}$, consider the function $f_t(x) := \mathbf{1}\{\sin(tx) \ge 0\}$, $x \in [-1,1]$. Prove that the function class $\{f_t : [-1,1] \to \mathbb{R} : t \in \mathbb{R}\}$ has infinite VC dimension (Note that this shows that VC dimension is not equivalent to the number of parameters in a function class).

7.2 Covering number bound for VC classes of sets

Theorem 7.8. *There exists a universal constant K such that for any VC class $\mathcal{C}$ of sets, any probability measure Q, any $r \geq 1$, and $0 < \epsilon < 1$,*

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq K\, V(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{rV(\mathcal{C})}. \tag{77}$$

Proof. See [van der Vaart and Wellner, 1996, Theorem 2.6.4]. □

In the following we will prove a slightly weaker version of the above result.

Theorem 7.9. *For any VC class $\mathcal{C}$ of sets, any $r \geq 1$, and $0 < \epsilon < 1$,*[60]

$$\sup_Q N(\epsilon, \mathcal{C}, L_r(Q)) \leq \left(\frac{c_1}{\epsilon}\right)^{rc_2V(\mathcal{C})} \tag{78}$$

Here c_1 and c_2 are universal positive constants and the supremum is over all probability measures Q on $\mathcal{X}$.

Proof. Fix $0 < \epsilon < 1$. Let $X_1, \ldots, X_n$ be i.i.d. Q. Let $m := D(\epsilon, \mathcal{C}, L_1(Q))$ be the ϵ-packing number for the collection $\mathcal{C}$ in the norm $L_1(Q)$. Thus, there exists $C_1, \ldots, C_m \in \mathcal{C}$ which satisfy

$$Q|\mathbf{1}_{C_i} - \mathbf{1}_{C_j}| = Q(C_i \triangle C_j) > \epsilon, \qquad i \neq j.$$

Let $\mathcal{F} := \{\mathbf{1}_C : C \in \mathcal{C}\}$. We consider this function class view point as it is sometimes more natural than working with the collection of sets $\mathcal{C}$. Note that, $\{f_i \equiv \mathbf{1}_{C_i}\}_{i=1}^m$ is a set of m ϵ-separated functions in $\mathcal{F}$ in the $L_1(Q)$-metric, as, for $i \neq j$,

$$\epsilon < \int |f_i - f_j| dQ = Q\{f_i \neq f_j\} = Q(C_i \triangle C_j) = \mathbb{P}[X_1 \in C_i \triangle C_j].$$

By the above, we have

$$\mathbb{P}[f_i(X_1) = f_j(X_1)] = 1 - \mathbb{P}[f_i(X_1) \neq f_j(X_1)] = 1 - \mathbb{P}[X_1 \in C_i \triangle C_j] < 1 - \epsilon \leq e^{-\epsilon}.$$

By the independence of $X_1, \ldots, X_n$ we deduce then that for every $k \geq 1$,

$$\mathbb{P}[f_i(X_1) = f_j(X_1), \ldots, f_i(X_k) = f_j(X_k)] \leq e^{-k\epsilon}.$$

In words, this means that the probability that f_i and f_j agree on every $X_1, \ldots, X_k$ is at most $e^{-k\epsilon}$. By the union bound, we have

$$\mathbb{P}\left[(f_i(X_1), \ldots, f_i(X_k)) = (f_j(X_1), \ldots, f_j(X_k)) \text{ for some } 1 \leq i < j \leq m\right] \leq \binom{m}{2} e^{-k\epsilon} \leq \frac{m^2}{2} e^{-k\epsilon}.$$

Recalling that $\mathcal{F}(x_1, \ldots, x_k) = \{(f(x_1), \ldots, f(x_k)) : f \in \mathcal{F}\}$, this immediately gives

$$\mathbb{P}[|\mathcal{F}(X_1, \ldots, X_k)| \geq m] \geq 1 - \frac{m^2}{2} e^{-k\epsilon}.$$

[60] Note that $N(\epsilon, \mathcal{C}, L_r(Q)) = 1$ for all $\epsilon \geq 1$.

Thus if we take $k := \left\lceil \frac{2\log m}{\epsilon} \right\rceil \geq \frac{2\log m}{\epsilon}$, then, $\mathbb{P}[|\mathcal{F}(X_1, \ldots, X_k)| \geq m] \geq 1/2$. Thus for the choice of k above, there exists a subset $\{z_1, \ldots, z_k\}$ of cardinality k such that $|\mathcal{F}(z_1, \ldots, z_k)| \geq m$. We now apply the Sauer-Shelah-VC lemma and deduce that

$$m \leq |\mathcal{F}(z_1, \ldots, z_k)| \leq \max_{x_1, \ldots, x_k} \Delta_k(\mathcal{C}; x_1, \ldots, x_k) \leq \sum_{j=1}^{V(\mathcal{C})} \binom{k}{j}. \tag{79}$$

We now split into two cases depending on whether $k \leq V(\mathcal{C})$ or $k \geq V(\mathcal{C})$.

Case 1: $k \leq V(\mathcal{C})$. Here (79) gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq 2^{V(\mathcal{C})} \leq \left(\frac{2}{\epsilon}\right)^{V(\mathcal{C})},$$

which proves (78).

Case 2: $k \geq V(\mathcal{C})$. Here (79) gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left(\frac{ke}{V(\mathcal{C})}\right)^{V(\mathcal{C})},$$

so that using the choice of k which satisfies $k \leq \frac{4\log m}{\epsilon}$,

$$m^{1/V(\mathcal{C})} \leq \frac{ke}{V(\mathcal{C})} \leq \frac{4e}{V(\mathcal{C})\epsilon} \log m = \frac{8e}{\epsilon} \log m^{1/(2V(\mathcal{C}))} \leq \frac{8e}{\epsilon} m^{1/(2V(\mathcal{C}))},$$

where we have used $\log x \leq x$. This immediately gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left(\frac{8e}{\epsilon}\right)^{2V(\mathcal{C})},$$

which completes the proof of the result for $r = 1$.

For $L_r(Q)$ with $r > 1$, note that

$$\|\mathbf{1}_C - \mathbf{1}_D\|_{L_1(Q)} = Q(C \triangle D) = \|\mathbf{1}_C - \mathbf{1}_D\|_{L_r(Q)}^r,$$

so that

$$N(\epsilon, \mathcal{C}, L_r(Q)) = N(\epsilon^r, \mathcal{C}, L_1(Q)) \leq \left(c_1 \epsilon^{-r}\right)^{c_2 V(\mathcal{C})}.$$

This completes the proof. □

Exercise (HW3): Suppose $\mathcal{F}$ is a Boolean class of functions with VC dimension $V(\mathcal{F})$. Then, for some constant $C > 0$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f|\right] \leq C\sqrt{\frac{V(\mathcal{F})}{n}}.$$

Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued observations having a common cdf F. Apply this result to obtain a high probability upper bound on $\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)|$, i.e., show that

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{C}{\sqrt{n}} + \sqrt{\frac{2}{n} \log \frac{1}{\alpha}}$$

with probability at least $1 - \alpha$ (for $\alpha \in (0, 1)$).

Example 7.10 (Classification). *Recall the problem of classification from Section 1.4 where we observe i.i.d. data $(Z_1, Y_1), \ldots, (Z_n, Y_n) \sim P$ with $Z_i \in \mathcal{Z}$ and $Y_i \in \{0, 1\}$. Let $\mathcal{C}$ be a class of functions from $\mathcal{Z}$ to $\{0, 1\}$ — the class of classifiers under consideration. The empirical risk minimizer classifier is $\hat{g}_n := \mathrm{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n I\{g(Z_i) \neq Y_i\}$. It is usually of interest to understand the test error of $\hat{g}_n$ relative to the best test error in the class $\mathcal{C}$, i.e., $L(\hat{g}_n) - \inf_{g \in \mathcal{C}} L(g)$ (here $L(g) := \mathbb{P}(g(Z) \neq Y)$ is the misclassification error of g). If g^* minimizes $L(g)$ over $g \in \mathcal{C}$, then we have seen in (5) that*

$$L(\hat{g}_n) - L(g^*) \leq 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)| = 2 \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P) f|$$

where

$$\mathcal{F} := \{(z, y) \mapsto I\{g(z) \neq y\} : g \in \mathcal{C}\}.$$

Using the bounded differences concentration inequality and the bound given by Example 4.8, we obtain

$$L(\hat{g}_n) - L(g^*) \leq C \sqrt{\frac{V(\mathcal{F})}{n}} + \sqrt{\frac{8}{n} \log \frac{1}{\alpha}}$$

with probability at least $1 - \alpha$ (for $\alpha \in (0, 1)$). The above display would be useful if we could upper bound $V(\mathcal{F})$ effectively. We will now show that $V(\mathcal{F}) \leq V(\mathcal{C})$. To see this, it is enough to argue that if $\mathcal{F}$ can shatter $(z_1, y_1), \ldots, (z_n, y_n)$, then $\mathcal{C}$ can shatter $z_1, \ldots, z_n$. For this, let $\eta_1, \ldots, \eta_n$ be arbitrary points in $\{0, 1\}$. We need to obtain a function $g \in \mathcal{C}$ such that $g(z_i) = \eta_i$, for $i = 1, \ldots, n$. Define $\delta_1, \ldots, \delta_n$ by

$$\delta_i := \eta_i I\{y_i = 0\} + (1 - \eta_i) I\{y_i = 1\}.$$

As $\mathcal{F}$ can shatter $(z_1, y_1), \ldots, (z_n, y_n)$, there exists $f \in \mathcal{F}$, say $f(z, y) = I\{g(z) \neq y\}$ for some $g \in \mathcal{C}$, with $f(z_i, y_i) = \delta_i$, for $i = 1, \ldots, n$. Then, $g(z_i) = \eta_i$[61], for $i = 1, \ldots, n$. This proves that $\mathcal{C}$ shatters $z_1, \ldots, z_n$ and completes the proof of the fact that $V(\mathcal{F}) \leq V(\mathcal{C})$. Thus, we obtain

$$L(\hat{g}_n) - L(g^*) \leq 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)| \leq C \sqrt{\frac{V(\mathcal{C})}{n}} + \sqrt{\frac{8}{n} \log \frac{1}{\alpha}}$$

with probability at least $1 - \alpha$. In fact, this is one of the important results in the VC theory.

7.3 VC classes of functions

Let us start with a motivating application. Recall from Example 1.5 the class of functions $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$ where $f_t(x) = |x - t|$. In Example 1.5 we needed to show asymptotic

[61] First observe that $f(z_i, y_i) = 0 \Leftrightarrow g(z_i) = y_i$ and $f(z_i, y_i) = 1 \Leftrightarrow g(z_i) \neq y_i$. Suppose that $\delta_i = 0$, i.e., $f(z_i, y_i) = 0$. Then, we must have, from the definition of δ_i, $0 = \eta_i I\{y_i = 0\} = (1 - \eta_i) I\{y_i = 1\}$, which implies that $y_i = \eta_i$ and thus, $g(z_i) = y_i = \eta_i$. Similarly, suppose that $\delta_i = 1$, i.e., $f(z_i, y_i) = 1$. Then, we must have, from the definition of δ_i, $y_i \neq \eta_i$, and thus as $g(z_i) \neq y_i$, we have $g(z_i) = \eta_i$. Thus, in both cases we see that $g(z_i) = \eta_i$.

equicontinuity of a certain process which boiled down to controlling the modulus of continuity of the empirical process indexed by $\mathcal{F}$ as in (48). In particular, we may ask: "Is this function class 'nice' is some sense so that results analogous to (80), and thus (48), hold?".

The *VC subgraph dimension* of $\mathcal{F}$ is simply the VC dimension of the Boolean class obtained by taking the indicators of the subgraphs of functions in $\mathcal{F}$. To formally define this, let us first define the notion of *subgraph* of a function.

Definition 7.11. *The subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is a subset of $\mathcal{X} \times \mathbb{R}$ defined as*

$$C_f := \{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\}.$$

A collection $\mathcal{F}$ of measurable functions on $\mathcal{X}$ is called a VC subgraph class, or just a VC class, if the collection of all subgraphs of the functions in $\mathcal{F}$ (i.e., $\{C_f : f \in \mathcal{F}\}$) forms a VC class of sets (in $\mathcal{X} \times \mathbb{R}$).

Let $V(\mathcal{F})$ be the VC dimension of the set of subgraphs of functions in $\mathcal{F}$. Just as for sets, the covering numbers of VC classes of functions grow at a polynomial rate.

Theorem 7.12. *For a VC class of functions $\mathcal{F}$ with measurable envelope function F and $r \geq 1$, one has for any probability measure Q with $\|F\|_{Q,r} > 0$,*

$$N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq K\, V(\mathcal{F})\, (4e)^{V(\mathcal{F})} \left(\frac{2}{\epsilon}\right)^{rV(\mathcal{F})}, \tag{80}$$

for a universal K and $0 < \epsilon < 1$.

Proof. Let $\mathcal{C}$ be the set of all subgraphs C_f of functions f in $\mathcal{F}$. Note that $Q|f-g| = (Q \times \lambda)(\mathbf{1}_{C_f \Delta C_g}) = (Q\times\lambda)|\mathbf{1}_{C_f} - \mathbf{1}_{C_g}|$ where λ is the Lebesgue measure on $\mathbb{R}$[62]. Renormalize $Q\times\lambda$ to a probability measure on the set $\{(x,t) : |t| \leq F(x)\}$ by defining $P = (Q \times \lambda)/(2\|F\|_{Q,1})$. Thus, as $P(C_f \Delta C_g) = P|\mathbf{1}_{C_f} - \mathbf{1}_{C_g}| = \frac{1}{2\|F\|_{Q,1}} Q|f-g|$,

$$N(\epsilon 2\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) = N(\epsilon, \mathcal{C}, L_1(P)).$$

Then by the result for VC classes of sets stated in Theorem 7.8,

$$N(\epsilon 2\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) \leq KV(\mathcal{F}) \left(\frac{4e}{\epsilon}\right)^{V(\mathcal{F})}, \tag{81}$$

for a universal constant K, for any probability measure Q with $\|F\|_{Q,1} > 0$. This completes the proof for $r = 1$.

For $r > 1$ note that

$$Q|f-g|^r \leq Q[|f-g|(2F)^{r-1}] = 2^{r-1} R[|f-g|] Q[F^{r-1}], \tag{82}$$

[62]Fact: For any two real numbers a and b, we have the identity $|a-b| = \int |I\{t < a\} - I\{t < b\}| dt$.

for the probability measure R with density $F^{r-1}/Q[F^{r-1}]$ with respect to Q. We claim that

$$N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq N(2(\epsilon/2)^r R[F], \mathcal{F}, L_1(R)). \tag{83}$$

To prove this claim let $N = N(2(\epsilon/2)^r R[F], \mathcal{F}, L_1(R))$ and let $f_1, \ldots f_N$ an $2(\epsilon/2)^r R[F]$-net for the class $\mathcal{F}$ (under $L_1(R)$-norm). Therefore, given $f \in \mathcal{F}$, there exists $k \in \{1, \ldots, N\}$ such that $\|f - f_k\|_{L_1(R)} \leq 2(\epsilon/2)^r R[F]$. Hence, by (82),

$$Q|f - f_k|^r \leq 2^{r-1} R[|f - f_k|]\, Q[F^{r-1}] \leq 2^{r-1} 2(\epsilon/2)^r R[F]\, Q[F^{r-1}] = \epsilon^r Q[F^r],$$

which implies that $\|f - f_k\|_{L_r(Q)} \leq \epsilon\|F\|_{Q,r}$. Thus, we have obtained a $\epsilon\|F\|_{Q,r}$-cover of $\mathcal{F}$ in the $L_r(Q)$-norm, which proves the claim. Now, combining (83) with (81) yields

$$N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq K V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{2}{\epsilon}\right)^{rV(\mathcal{F})},$$

which completes the proof. □

The preceding theorem shows that a VC class has a finite uniform entropy integral, with much to spare. In fact we can show that if $\mathcal{F}$ is a class of measurable functions with envelope F and VC subgraph dimension $V(\mathcal{F})$ then the expected supremum of the empirical process can be easily controlled[63].

[63] Here in a maximal inequality for an important (VC) class of functions we will encounter soon.

Theorem 7.13. *Let $\mathcal{F}$ be a measurable class of functions with a constant envelope U such that for $A > e^2$ and $V \geq 2$ and for every finitely supported probability measure Q*

$$N(\epsilon U, \mathcal{F}, L_2(Q)) \leq \left(\frac{A}{\epsilon}\right)^V, \qquad 0 \leq \epsilon < 1.$$

Then, for all n,

$$\mathbb{E}\Big\|\sum_{i=1}^n (f(X_i) - Pf)\Big\|_{\mathcal{F}} \leq L\left(\sqrt{n}\sigma\sqrt{V \log\frac{AU}{\sigma}} \vee VU \log\frac{AU}{\sigma}\right) \tag{84}$$

where L is a universal constant and σ is such that $\sup_{f\in\mathcal{F}} P(f - Pf)^2 \leq \sigma^2$.

Remark 7.2. *If $n\sigma^2 \gtrsim V\log(AU/\sigma)$ then the above result shows that*

$$\mathbb{E}\Big\|\sum_{i=1}^n (f(X_i) - Pf)\Big\|_{\mathcal{F}} \lesssim \sqrt{n\sigma^2 V \log(AU/\sigma)},$$

which means that if $n\sigma^2$ is not too small, then the 'price' one pays for considering the expectation of the supremum of infinitely many sums instead of just one is the factor $\sqrt{V\log(AU/\sigma)}$.

Proof. We assume without loss of generality that the class $\mathcal{F}$ contains the function 0 and that the functions f are P-centered. It suffices to prove the inequality in the theorem for $U = 1$. By our symmetrization result (see Theorem 3.17), we have $\mathbb{E}\|\sum_{i=1}^n f(X_i)\|_{\mathcal{F}} \leq 2\mathbb{E}\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}$. By Dudley's entropy bound (see Theorem 4.1 and (41)), we have

$$\frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\Big\|\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}} \leq K\sqrt{V}\int_0^{\sqrt{\|\mathbb{P}_n f^2\|_{\mathcal{F}}}} \sqrt{\log\frac{A}{\epsilon}}\, d\epsilon,$$

Exercise (HW3): Suppose $\mathcal{F}$ is a class of measurable functions with envelope F and VC subgraph dimension $V(\mathcal{F})$. Then, for some constant $C > 0$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f|\right] \le C\|F\|_{P,2}\sqrt{\frac{V(\mathcal{F})}{n}}. \tag{85}$$

Exercise (HW3): Recall the setting of Example 5.5 which considered a change point problem. Show that condition (45) (bound on the modulus of continuity of the empirical process) needed to apply Theorem 5.2, to obtain the rate of convergence of the estimator, holds with an appropriate function $\phi_n(\cdot)$ (also see Remark 5.2).

---

where ε_i's are i.i.d. Rademacher variables independent of the variables X_j's and $\mathbb{E}_\varepsilon$ indicates expectation with respect to the ε_i's. It is easy to see that if $\log(C/c) \ge 2$ then

$$\int_0^c \left(\log\frac{C}{x}\right)^{1/2} dx \le 2c\left(\log\frac{C}{c}\right)^{1/2}.$$

Since $A/\|\mathbb{P}_n f^2\|_{\mathcal{F}} \ge e^2$ (as $|f| \le 1$ by assumption), we conclude that

$$\frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\Big\|\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}} \le 2K\sqrt{V}\sqrt{\|\mathbb{P}_n f^2\|_{\mathcal{F}}}\sqrt{\log\frac{A}{\sqrt{\|\mathbb{P}_n f^2\|_{\mathcal{F}}}}}.$$

By the concavity of the function $\sqrt{x(-\log x)}$ on $(0, e^{-1})$, this yields

$$\frac{1}{\sqrt{n}}\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}} \le 2K\sqrt{V}\sqrt{\mathbb{E}\|P_n f^2\|_{\mathcal{F}}}\sqrt{\log\frac{A}{\sqrt{\mathbb{E}\|P_n f^2\|_{\mathcal{F}}}}} =: 2K\sqrt{V}B.$$

Next, notice that

$$\mathbb{E}\|\mathbb{P}_n f^2\|_{\mathcal{F}} \le \sigma^2 + \frac{1}{n}\mathbb{E}\Big\|\sum_{i=1}^n (f^2(X_i) - Pf^2)\Big\|_{\mathcal{F}} \le \sigma^2 + \frac{2}{n}\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i f^2(X_i)\Big\|_{\mathcal{F}}.$$

Now, since $\mathbb{P}_n[(f^2 - g^2)^2] \le 4\mathbb{P}_n[(f-g)^2]$ (as $|f| \le 1$ by assumption), which implies $N(\epsilon, \mathcal{F}^2, L_2(\mathbb{P}_n)) \le N(\epsilon/2, \mathcal{F}, L_2(\mathbb{P}_n))$, we can estimate the last expectation again by the entropy bound (2.14), and get, with the change of variables $\epsilon/2 = u$,

$$\mathbb{E}\|\mathbb{P}_n f^2\|_{\mathcal{F}} \le \sigma^2 + \frac{4K\sqrt{V}}{\sqrt{n}}\mathbb{E}\left[\int_0^{\sqrt{\|\mathbb{P}_n f^2\|_{\mathcal{F}}}} \sqrt{\log\frac{A}{u}}\, du\right]$$

which, by the previous computations gives

$$\mathbb{E}\|\mathbb{P}_n f^2\|_{\mathcal{F}} \le \sigma^2 + \frac{8K\sqrt{V}B}{\sqrt{n}}$$

Replacing this into the definition of B shows that B satisfies the inequality (check!)

$$B^2 \le \left(\sigma^2 + \frac{8K\sqrt{V}B}{\sqrt{n}}\right)\log\frac{A}{\sigma}.$$

This implies that B is dominated by the largest root of the quadratic function $B^2 - 8K\sqrt{V}B\log(A/\sigma)/\sqrt{n} - \sigma^2\log(A/\sigma)$ which yields the desired result for a suitable constant L. ☐

Note that when the dominant term is the first, we only pay a logarithmic price for the fact that we are taking the supremum over a countable or uncountable set. In fact the inequality is sharp in the range of (σ, n) where the first term dominates.

7.4 Examples and Permanence Properties

The results of this subsection give basic methods for generating VC (subgraph) classes. This is followed by a discussion of methods that allow one to build up new function classes related to the VC property (from basic classes).

Although it is obvious, it is worth mentioning that a subclass of a VC class is itself a VC class. The following lemma shows that various operations on VC classes (of sets) preserve the VC structure.

Lemma 7.14. *Let $\mathcal{C}$ and $\mathcal{D}$ be VC classes of sets in a set $\mathcal{X}$ and $\phi : \mathcal{X} \to \mathcal{Y}$ and $\psi : \mathcal{Z} \to \mathcal{X}$ be fixed functions. Then:*

(i) $\mathcal{C}^c = \{C^c : C \in \mathcal{C}\}$ *is VC;*

(ii) $\mathcal{C} \sqcap \mathcal{D} = \{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$ *is VC;*

(iii) $\mathcal{C} \sqcup \mathcal{D} = \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ *is VC;*

(iv) $\phi(\mathcal{C})$ *is VC if ϕ is one-to-one;*

(v) $\psi^{-1}(\mathcal{C})$ *is VC;*

(vi) $\mathcal{C} \times \mathcal{D} = \{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}$ *is VC in $\mathcal{X} \times \mathcal{Y}$.*

Proof. (i) The set C^c picks out the points of a given set $\{x_1, \ldots, x_n\}$ that C does not pick out. Thus if $\mathcal{C}$ shatters a given set of points, so does $\mathcal{C}^c$. This proves (i) and shows that the dimensions of $\mathcal{C}$ and $\mathcal{C}^c$ are equal.

(ii) Fix $n \geq \max\{V(\mathcal{C}), V(\mathcal{D})\}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ be arbitrary. We have to study the cardinality of the set $\{C \cap D \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}, D \in \mathcal{D}\}$. From the n points $x_1, \ldots, x_n$, $\mathcal{C}$ can pick out $O(n^{V(\mathcal{C})})$ subsets. From each of these subsets, $\mathcal{D}$ can pick out at most $O(n^{V(\mathcal{D})})$ further subsets[64]. Thus $\mathcal{C} \sqcap \mathcal{D}$ can pick out $O(n^{V(\mathcal{C})+V(\mathcal{D})})$ subsets. For large n, this is certainly smaller than 2^n. This proves (ii).

Next, (iii) follows from a combination of (i) and (ii), since $C \cup D = (C^c \cap D^c)^c$.

(iv) We will show that if $\phi(\mathcal{C})$ shatters a set of points in $\mathcal{Y}$ then $\mathcal{C}$ should also shatter a set of points of the same cardinality in $\mathcal{X}$, which will yield the desired result. Suppose that $\phi(\mathcal{C})$ shatters $\{y_1, \ldots, y_n\} \subset \mathcal{Y}$. Then each y_i must be in the range of ϕ and there exist $x_1, \ldots, x_n$ such that ϕ is a bijection between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. We now claim that $\mathcal{C}$ must shatter $\{x_1, \ldots, x_n\}$. To see this, let $A := \{x_{i_1}, \ldots, x_{i_k}\}$ for $1 \leq i_1, \ldots, i_k \leq n$ distinct, with $0 \leq k \leq n$. As $\phi(\mathcal{C})$ shatters $\{y_1, \ldots, y_n\} \subset \mathcal{Y}$, $\phi(\mathcal{C})$ picks out there $\{y_{i_1}, \ldots, y_{i_k}\}$ and thus there exists $B \equiv \phi(C) \in \phi(\mathcal{C})$, where $C \in \mathcal{C}$, such that $B \cap \{y_1, \ldots, y_n\} = \{y_{i_1}, \ldots, y_{i_k}\}$.

[64] Note that by Sauer's lemma, for any $1 \leq k \leq n$, $\Delta_n(\mathcal{D}; x_{i_1}, \ldots, x_{i_k}) \leq \sum_{j=0}^{V(\mathcal{D})} \binom{k}{j} \leq O(n^{V(\mathcal{D})})$, for any $\{x_{i_1}, \ldots, x_{i_k}\} \subset \{x_1, \ldots, x_n\}$.

As ϕ is a bijection, this means that $C \cap \{x_1, \ldots, x_n\} = \{x_{i_1}, \ldots, x_{i_k}\}$, and thus $\mathcal{C}$ picks out there $\{x_{i_1}, \ldots, x_{i_k}\}$.

For (v) the argument is analogous: if $\psi^{-1}(\mathcal{C})$ shatters $z_1, \ldots, z_n$, then all $x_i := \psi(z_i)$ must be different and the restriction of ψ to $z_1, \ldots, z_n$ is a bijection on its range. We now claim that then $\mathcal{C}$ shatters $x_1, \ldots, x_n$[65]. Thus, as $\mathcal{C}$ has finite VC dimension, so has $\psi^{-1}(\mathcal{C})$.

For (vi) note first that $\mathcal{C} \times \mathcal{Y}$ and $\mathcal{X} \times \mathcal{D}$ are VC classes[66]. Then by (ii) so is their intersection $\mathcal{C} \times \mathcal{D}$. $\square$

Exercise (HW3) (Open and closed subgraphs): For a set $\mathcal{F}$ of measurable functions, define "closed" and "open" subgraphs by $\{(x,t) : t \le f(x)\}$ and $\{(x,t) : t < f(x)\}$, respectively. Then the collection of "closed" subgraphs has the same VC-dimension as the collection of "open" subgraphs. Consequently, "closed" and "open" are equivalent in the definition of a VC-subgraph class.

Lemma 7.15. *Any finite-dimensional vector space $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is VC subgraph of dimension smaller than or equal to dim$(\mathcal{F}) + 1$.*

Proof. By assumption, there exists $m := \dim(\mathcal{F})$ functions $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$ such that

$$\mathcal{F} := \Big\{ \sum_{i=1}^{m} \alpha_j f_j(x) : \alpha_j \in \mathbb{R} \Big\}.$$

Take any collection of $n = \dim(\mathcal{F}) + 2$ points $(x_1, t_1), \ldots, (x_n, t_n)$ in $\mathcal{X} \times \mathbb{R}$. We will show that the subgraphs of $\mathcal{F}$ do not shatter these n points. Let $H \in \mathbb{R}^{n \times m}$ be the matrix with elements $(f_j(x_i))$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. By assumption, the vectors in

$$\mathcal{H} := \{(f(x_1) - t_1, \ldots, f(x_n) - t_n) : f \in \mathcal{F}\} = \{H\mathbf{c} - (t_1, \ldots, t_n) : \mathbf{c} \in \mathbb{R}^m\},$$

are contained in a $\dim(\mathcal{F}) + 1 = (n-1)$-dimensional subspace[67] of $\mathbb{R}^n$. Any vector $a \neq 0$ that is orthogonal to this subspace satisfies

$$\sum_{i : a_i > 0} a_i (f(x_i) - t_i) = \sum_{i : a_i \le 0} (-a_i)(f(x_i) - t_i), \qquad \text{for every } f \in \mathcal{F}.$$

(Define the sum over an empty set as zero.) There exists such a vector a with at least one strictly positive coordinate. We will show that the subgraphs of $\mathcal{F}$ do not pick out the set $\{(x_i, t_i) : a_i > 0\}$.

Suppose there exists $f \in \mathcal{F}$ such that $C_f \cap \{(x_i, t_i)\}_{i=1}^n = \{(x_i, t_i) : a_i > 0\}$. Then, for i such that $a_i \le 0$ we must have $(x_i, t_i) \notin C_f$, i.e., $t_i \ge f(x_i)$. However, then the left side of the above display would be strictly positive and the right side non-positive for this

[65] Exercise (HW3): Show this.

[66] Exercise (HW3): Show this.

[67] We can find $\alpha \in \mathbb{R}^n$ such that $\alpha^\top H = 0$ and $\alpha^\top (t_1, \ldots, t_n) = 0$. Such an α exists as the columns of H and $(t_1, \ldots, t_n)$ span at most an $m + 1 = n - 1$ dimensional subspace of $\mathbb{R}^n$.

f. Conclude that the subgraphs of $\mathcal{F}$ do not pick out the set $\{(x_i, t_i) : a_i > 0\}$. Hence the subgraphs of $\mathcal{F}$ shatter no set of n points. $\square$

Example 7.16. *Let $\mathcal{F}$ be the set of all linear combinations $\sum \lambda_i f_i$ of a given, finite set of functions $f_1, \ldots, f_k$ on $\mathcal{X}$. Then $\mathcal{F}$ is a VC class and hence has a finite uniform entropy integral. Furthermore, the same is true for the class of all sets $\{f > c\}$ if f ranges over $\mathcal{F}$ and c over $\mathbb{R}$.*

Lemma 7.17. *The set of all translates $\{\psi(x - h) : h \in \mathbb{R}\}$ of a fixed monotone function $\psi : \mathbb{R} \to \mathbb{R}$ is VC of dimension 1.*

Proof. By the monotonicity, the subgraphs are linearly ordered by inclusion: if ψ is nondecreasing, then the subgraph of $x \mapsto \psi(x-h_1)$ is contained in the subgraph of $x \mapsto \psi(x-h_2)$ if $h_1 \geq h_2$. Any collection of sets with this property has VC dimension 1 by Proposition 7.18[68]. $\square$

Lemma 7.19. *Let $\mathcal{F}$ and $\mathcal{G}$ be VC subgraph classes of functions on a set $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}, \phi : \mathbb{R} \to \mathbb{R}$, and $\psi : \mathcal{Z} \to \mathcal{X}$ fixed functions. Then,*

(i) $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g : f \in \mathcal{F}; g \in \mathcal{G}\}$ *is VC subgraph;*

(ii) $\mathcal{F} \vee \mathcal{G}$ *is VC subgraph;*

(iii) $\{\mathcal{F} > 0\} := \{\{f > 0\} : f \in \mathcal{F}\}$ *is VC;*

(iv) $-\mathcal{F}$ *is VC;*

(v) $\mathcal{F} + g := \{f + g : f \in \mathcal{F}\}$ *is VC subgraph;*

(vi) $\mathcal{F} \cdot g = \{fg : f \in \mathcal{F}\}$ *is VC subgraph;*

(vii) $\mathcal{F} \circ \psi = \{f(\psi) : f \in \mathcal{F}\}$ *is VC subgraph;*

(viii) $\phi \circ \mathcal{F}$ *is VC subgraph for monotone ϕ.*

68

Proposition 7.18. *Suppose that $\mathcal{C}$ is a collection of at least two subsets of a set $\mathcal{X}$. Show that $V(\mathcal{C}) = 1$ if either (a) $\mathcal{C}$ is linearly ordered by inclusion, or, (b) any two sets in $\mathcal{C}$ are disjoint.*

Proof. Consider (a) first. Take points $x_1, x_2 \in \mathcal{X}$. We need to show that this set of 2 points cannot be shattered. Suppose that $\{x_1, x_2\}$ can be shattered. Let C_1 pick out $\{x_1\}$ and C_2 pick out $\{x_2\}$. By (a), one of these sets is contained in the other. Suppose $C_1 \subset C_2$. But then $\{x_1\} \subset C_2$ and this contradicts the fact that C_2 picks out $\{x_2\}$. On the other hand, at least one set of size 1 is shattered.

Next consider (b). As before, suppose that $\{x_1, x_2\}$ can be shattered. Suppose C picks out $\{x_1\}$ and D picks out $\{x_1, x_2\}$. But then C and D are no longer disjoint. $\square$

Proof. The subgraphs of $f \wedge g$ and $f \vee g$ are the intersection and union of the subgraphs of f and g, respectively. Hence (i) and (ii) are consequences of Lemma 7.14.

For (iii) note that the sets $\{f > 0\}$ are one-to-one images of the intersections of the (open) subgraphs with the set $\mathcal{X} \times \{0\}$, i.e.,

$$\{f > 0\} = \{x \in \mathcal{X} : f(x) > 0\} = \phi\big(\{(x,t) \in \mathcal{X} \times \mathbb{R} : f(x) > t\} \cap (\mathcal{X} \times \{0\})\big).$$

Here $\phi : \mathcal{X} \times \{0\} \to \mathcal{X}$ defined as $\phi(x, 0) = x$ is one-one. Thus the class $\{\mathcal{F} > 0\}$ is VC by (ii) and (iv) of Lemma 7.14.

(iv) The subgraphs of the class $-\mathcal{F}$ are the images of the open supergraphs of $\mathcal{F}$ under the map $(x,t) \mapsto (x, -t)$. The open supergraphs are the complements of the closed subgraphs, which are VC by the previous exercise. Now (iv) follows from the previous lemma.

For (v) it suffices to note that the subgraphs of the class $\mathcal{F} + g$ shatter a given set of points $(x_1, t_1), \ldots, (x_n, t_n)$ if and only if the subgraphs of $\mathcal{F}$ shatter the set $(x_i, t_i - g(x_i))$.

The subgraph of the function fg is the union of the sets

$$\begin{aligned}
C^+ &:= \{(x,t) : t < f(x)g(x), g(x) > 0\},\\
C^- &:= \{(x,t) : t < f(x)g(x), g(x) < 0\},\\
C^0 &:= \{(x,t) : t < 0, g(x) = 0\},
\end{aligned}$$

It suffices to show that these sets are VC in $(\mathcal{X} \cap \{g > 0\}) \times \mathbb{R}$, $(\mathcal{X} \cap \{g < 0\}) \times \mathbb{R}$, and $(\mathcal{X} \cap \{g = 0\}) \times \mathbb{R}$, respectively[69]. Now, for instance, $\{i : (x_i, t_i) \in C^-\}$ is the set of indices of the points $(x_i, t_i/g(x_i))$ picked out by the open supergraphs of $\mathcal{F}$. These are the complements of the closed subgraphs and hence form a VC class.

The subgraphs of the class $\mathcal{F} \circ \psi$ are the inverse images of the subgraphs of functions in $\mathcal{F}$ under the map $(z,t) \mapsto (\psi(z), t)$. Thus (v) of Lemma (7.14) implies (vii).

For (viii) suppose that the subgraphs of $\phi \circ \mathcal{F}$ shatter the set of points $(x_1, t_1), \ldots, (x_n, t_n)$. Choose $f_1, \ldots, f_m$ from $\mathcal{F}$ such that the subgraphs of the functions $\phi \circ f_j$ pick out all $m = 2^n$ subsets. For each fixed i, define $s_i = \max\{f_j(x_i) : \phi(f_j(x_i)) \le t_i\}$. Then $s_i < f_j(x_i)$ if and only if $t_i < \phi(f_j(x_i))$, for every i and j, and the subgraphs of $f_1, \ldots, f_m$ shatter the points (x_i, s_i). $\square$

Exercise (HW3): The class of all ellipsoids $\{x \in \mathbb{R}^d : (x - \mu)^\top A(x - \mu) \le c\}$, for μ ranging over $\mathbb{R}^d$ and A ranging over the nonnegative $d \times d$ matrices, is VC. [**Hint**: This follows by a combination of Lemma 7.19(iii), Lemma 7.14(i) and Lemma 7.15. The third shows that the set of functions $x \mapsto (x - \mu)^\top A(x - \mu) - c$ (a vector space with basis functions $x \mapsto c, x \mapsto x_i$ and $x \mapsto x_i x_j$) is VC, and the first and second show that their positive (or negative) sets are also VC.

[69] Exercise (HW3): If $\mathcal{X}$ is the union of finitely many disjoint sets $\mathcal{X}_i$, and $\mathcal{C}_i$ is a VC class of subsets of $\mathcal{X}_i$ for each i, $i = 1, \ldots, m$, then $\sqcup_{i=1}^m \mathcal{C}_i$ is a VC class in $\cup_{i=1}^m \mathcal{X}_i$ of dimension $\sum_{i=1}^m V(\mathcal{C}_i)$.

Example 7.20. *Let $f_t(x) = |x - t|$ where $t \in \mathbb{R}$. Let $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$. Then this is a VC class of functions as $f_t(x) = (x - t) \vee [-(x - t)]$ and we can use Lemma 7.15 with Lemma 7.19(i) to prove the result.*

Sometimes, the result of simple operations on VC classes of functions (e.g., addition of two such classes) can produce a function classes that is not itself VC. However, the resulting function class might still have a uniform polynomial bound on covering numbers (as in (80)) and hence are very easy to work with. The following results provide a few such examples.

Lemma 7.21. *Fix $r \geq 1$. Suppose that $\mathcal{F}$ and $\mathcal{G}$ are classes of measurable functions with envelopes F and G respectively. Then, for every $0 < \epsilon < 1$,*

(i) $N(2\epsilon\|F + G\|_{Q,r}, \mathcal{F} + \mathcal{G}, L_r(Q)) \leq N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \cdot N(\epsilon\|G\|_{Q,r}, \mathcal{G}, L_r(Q))$;

(ii) $\sup_Q N(2\epsilon\|F\cdot G\|_{Q,2}, \mathcal{F}\cdot\mathcal{G}, L_2(Q)) \leq \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \ \sup_Q N(\epsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q))$, *where $\mathcal{F}\cdot\mathcal{G} := \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$, and the supremums are all taken over the appropriate subsets of all finitely discrete probability measures.*

Proof. Let us first prove (i). Find functions $f_1, \ldots, f_n$ and $g_1, \ldots, g_m$ such that

$$\min_i \|f - f_i\|_{Q,r}^r \leq \epsilon^r \|F\|_{Q,r}^r, \ \forall f \in \mathcal{F}, \ \text{ and } \ \min_j \|g - g_j\|_{Q,r}^r \leq \epsilon^r \|G\|_{Q,r}^r, \ \forall g \in \mathcal{G}.$$

Now, given $f + g \in \mathcal{F} + \mathcal{G}$, we can find i and j such that

$$\|f + g - f_i - g_j\|_{Q,r} \leq \|f - f_i\|_{Q,r} + \|g - g_j\|_{Q,r} \leq \epsilon\|F\|_{Q,r} + \epsilon\|G\|_{Q,r} \leq 2\epsilon\|F + G\|_{Q,r},$$

which completes the proof.

Let us prove (ii) now. Fix $\epsilon > 0$ and a finitely discrete probability measure $\tilde{Q}$ with $\|FG\|_{\tilde{Q},2} > 0$ (which also implies that $\|G\|_{\tilde{Q},2} > 0$), and let $dQ^* := G^2 d\tilde{Q}/\|G\|_{\tilde{Q},2}^2$. Clearly, Q^* is a finitely discrete probability measure with $\|F\|_{Q^*,2} > 0$. Let $f_1, f_2 \in \mathcal{F}$ satisfying $\|f_1 - f_2\|_{Q^*,2} \leq \epsilon\|F\|_{Q^*,2}$. Then

$$\epsilon \geq \frac{\|f_1 - f_2\|_{Q^*,2}}{\|F\|_{Q^*,2}} = \frac{\|(f_1 - f_2)G\|_{\tilde{Q},2}}{\|FG\|_{\tilde{Q},2}},$$

and thus, if we let $\mathcal{F}\cdot G := \{fG : f \in \mathcal{F}\}$,

$$N(\epsilon\|FG\|_{\tilde{Q},2}, \mathcal{F}\cdot G, L_2(\tilde{Q})) \leq N(\epsilon\|F\|_{Q^*,2}, \mathcal{F}, L_2(Q^*)) \leq \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)),$$

where the supremum is taken over all finitely discrete probability measures Q for which $\|F\|_{Q,2} > 0$. Since the right hand-side of the above display does not depend on $\tilde{Q}$, and since $\tilde{Q}$ satisfies $\|FG\|_{\tilde{Q},2} > 0$ but is otherwise arbitrary, we have that

$$\sup_Q N(\epsilon\|FG\|_{Q,2}, \mathcal{F}\cdot G, L_2(Q)) \leq \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)), \tag{86}$$

where the supremums are taken over all finitely discrete probability measures Q but with the left side taken over the subset for which $\|FG\|_{Q,2} > 0$ while the right side is taken over the subset for which $\|F\|_{Q,2} > 0$.

We can similarly show that the uniform entropy numbers for the class $\mathcal{G} \cdot F$ with envelope FG is bounded by the uniform entropy numbers for $\mathcal{G}$ with envelope G. Since $|f_1 g_1 - f_2 g_2| \leq |f_1 - f_2|G + |g_1 - g_2|F$ for all $f_1, f_2 \in \mathcal{F}$ and $g_1, g_2 \in \mathcal{G}$, part (i) in conjunction with (86) imply that

$$\sup_Q N(2\epsilon\|FG\|_{Q,2}, \mathcal{F}\cdot\mathcal{G}, L_2(Q)) \leq \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \times \sup_Q N(\epsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q)),$$

where the supremums are all taken over the appropriate subsets of all finitely discrete probability measures. $\square$

Lemma 7.22. *(i) Let $\rho(\cdot)$ be a real-valued right continuous function of bounded variation on $\mathbb{R}_+$. The covering number of the class $\mathcal{F}$ of all functions on $\mathbb{R}^d$ of the form $x \mapsto \rho(\|Ax + b\|)$, with A ranging over all $m \times d$ matrices and $b \in \mathbb{R}^m$ satisfies the bound*

$$N(\epsilon, \mathcal{F}, L_r(Q)) \leq K_1 \epsilon^{-V_1}, \tag{87}$$

for some K_1 and V_1 and for a constant envelope.

(ii) (Exercise (HW3)) Let $\lambda(\cdot)$ be a real-valued function of bounded variation on $\mathbb{R}$. The class of all functions on $\mathbb{R}^d$ of the form $x \mapsto \lambda(\alpha^\top x + \beta)$, with α ranging over $\mathbb{R}^d$ and β ranging over $\mathbb{R}$, satisfies (87) for a constant envelope.

Proof. Let us prove (i). By Lemma 7.21 it is enough to treat the two monotone components of $\rho(\cdot)$ separately. Assume, without loss of generality, that $\rho(\cdot)$ is bounded and nondecreasing, with $\rho(0) = 0$. Define $\rho^{-1}(\cdot)$ as the usual left continuous inverse of ρ on the range $T = (0, \sup \rho)$, i.e.,

$$\rho^{-1}(t) := \inf\{y : \rho(y) \geq t\}.$$

This definition ensures that

$$\{y : \rho(y) \geq t\} = \{y : y \geq \rho^{-1}(t)\}, \qquad \text{for } t \in T.$$

Exercise (HW3): Complete the proof now. $\square$

7.5 Exponential tail bounds: some useful inequalities

Suppose that we have i.i.d. data $X_1, \ldots, X_n$ on a set $\mathcal{X}$ having distribution P and $\mathcal{F}$ is a VC class of measurable real-valued functions on $\mathcal{X}$. We end this chapter with a brief discussion of some useful and historically important results on exponential tail bounds for the empirical process indexed by VC classes of functions. One of the classical results in this direction is the exponential tail bounds for the supremum distance between the empirical distribution and the true distribution function; see [Dvoretzky et al., 1956].

(A) **Empirical d.f., $\mathcal{X} = \mathbb{R}$**: Suppose that we consider the classical empirical d.f. of real-valued random variables. Thus, $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$. Then, letting $\mathbb{F}_n$ and F denote the empirical and true distribution functions, [Dvoretzky et al., 1956] showed that

$$\mathbb{P}(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \geq x) \leq C \exp(-2x^2)$$

for all $n \geq 1$, $x \geq 0$ where C is an absolute constant. [Massart, 1990] showed that $C = 2$ works, confirming a long-standing conjecture of Z. W. Birnbaum.

This result strengthens the GC theorem by quantifying the rate of convergence as n tends to infinity. It also estimates the tail probability of the Kolmogorov-Smirnov statistic.

(B) **Empirical d.f., $\mathcal{X} = \mathbb{R}^d$**: Now consider the classical empirical d.f. of i.i.d. random vectors: Thus $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}^d\}$. Then [Kiefer, 1961] showed that for every $\epsilon > 0$ there exists a C_ϵ such that

$$\mathbb{P}(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \geq x) \leq C_\epsilon \exp(-(2-\epsilon)x^2)$$

for all $n \geq 1$, $x > 0$.

(C) **Empirical measure, $\mathcal{X}$ general**: Let $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$ be such that

$$\sup_Q N(\epsilon, \mathcal{F}, L_1(Q)) \leq \left(\frac{K}{\epsilon}\right)^V,$$

where we assume that $V \geq 1$ and $K \geq 1$. Then [Talagrand, 1994] proved that

$$\mathbb{P}\Big(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{C}} \geq x\Big) \leq \frac{D}{x}\left(\frac{Dx^2}{V}\right)^V \exp(-2x^2) \tag{88}$$

for all $n \geq 1$ and $x > 0$, where $D \equiv D(K)$ depends on K only.

(D) **Empirical measure, $\mathcal{X}$ general**: Let $\mathcal{F}$ be a class of functions such that $f : \mathcal{X} \to [0,1]$ for every $f \in \mathcal{F}$, and $\mathcal{F}$ satisfies

$$\sup_Q N(\epsilon, \mathcal{F}, L_2(Q)) \leq \left(\frac{K}{\epsilon}\right)^V;$$

e.g., when $\mathcal{F}$ is a VC class $V = 2V(\mathcal{F})$. Then [Talagrand, 1994] proved that

$$\mathbb{P}\Big(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \geq x\Big) \leq \left(\frac{Dx}{\sqrt{V}}\right)^V \exp(-2x^2)$$

for all $n \geq 1$ and $x > 0$.

Example 7.23 (Projection pursuit)**.** *Projection pursuit (PP) is a type of statistical technique which involves finding the most "interesting" possible projections in multidimensional* *data. Often, projections which deviate more from a normal distribution are considered to be more interesting. The idea of projection pursuit is to locate the projection or projections from high-dimensional space to low-dimensional space that reveal the most details about the structure of the data set.*

Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. P on $\mathbb{R}^d$. The first step in PP is to estimate the distribution of the low-dimensional projections. We address this question here. In particular, we ask: "How large can d be (with n) so that we can still uniformly approximate all the one-dimensional projections of P?"

We answer the above questions below. For $t \in \mathbb{R}$ and $\gamma \in S^{d-1}$ (S^{d-1} is the unit sphere in $\mathbb{R}^d$), let

$$\mathbb{F}_n(t;\gamma) = \mathbb{P}_n[\mathbf{1}_{(-\infty,t]}(\gamma \cdot X)] = \mathbb{P}_n(\gamma \cdot X \leq t),$$

denote the empirical distribution function of $\gamma \cdot X_1, \ldots, \gamma \cdot X_n$. Let

$$F(t;\gamma) = P[\mathbf{1}_{(-\infty,t]}(\gamma \cdot X)] = \mathbb{P}(\gamma \cdot X_1 \leq t).$$

Question: *Under what conditions on $d = d_n \to \infty$ as $n \to \infty$, do we have*

$$D_n := \sup_{t \in \mathbb{R}} \sup_{\gamma \in S^{d-1}} |\mathbb{F}_n(t;\gamma) - F(t;\gamma)| \xrightarrow{\mathbb{P}} 0?$$

First note that the sets in question in this example are half-spaces

$$H_{t,\gamma} := \{x \in \mathbb{R}^d : \gamma \cdot x \leq t\}.$$

Note that

$$D_n = \sup_{t \in \mathbb{R}} \sup_{\gamma \in S^{d-1}} |\mathbb{P}_n(H_{t,\gamma}) - P(H_{t,\gamma})| = \|\mathbb{P}_n - P\|_{\mathcal{H}},$$

where $\mathcal{H} := \{H_{t,\gamma} : t \in \mathbb{R}, \gamma \in S^{d-1}\}$.

The key to answering the question raised in this example is one of the exponential bounds applied to the collection $\mathcal{H}$, the half-spaces in $\mathbb{R}^d$. The collection $\mathcal{H}$ is a VC collection of sets with $V(\mathcal{H}) = d + 1$[70].

[70] We have to prove two inequalities: $V(\mathcal{H}) \geq d+1$ and $V(\mathcal{H}) \leq d+1$. To prove the first inequality, we need to exhibit a particular set of size $d+1$ that is shattered by $\mathcal{H}$. Proving the second inequality is a bit more tricky: we need to show that for all sets of size $d+2$, there is labelling that cannot be realized using half-spaces.

Let us first prove $V(\mathcal{H}) \geq d+1$. Consider the set $\mathcal{X}_0 = \{0, e_1, \ldots, e_d\}$ which consists of the origin along with the vectors in the standard basis of $\mathbb{R}^d$ (also let $e_0 = 0 \in \mathbb{R}^d$). Let $A := \{e_{i_1}, \ldots, e_{i_m}\}$ be a subset of $\mathcal{X}_0$, where $m \geq 0$ and $\{i_1, \ldots, i_m\} \subseteq \{0, \ldots, d\}$. We will show that A is picked out by $\mathcal{H}$. Let $\gamma = (\gamma_1, \ldots, \gamma_d) \in S^{d-1}$ be such that $\gamma_j = -1/\sqrt{d}$ if $j \in \{i_1, \ldots, i_m\}$ and $\gamma_j = 1/\sqrt{d}$ otherwise. Let $t = 0$ if $e_0 \in A$ and $t = -1/\sqrt{d}$ if $e_0 \notin A$. Thus, for $x \in A$, $\gamma \cdot x \leq t$ and for $x \in \mathcal{X}_0 \setminus A$, $\gamma \cdot x > t$. Therefore, $\{x \in \mathbb{R}^d : \gamma \cdot x \leq t\} \cap \mathcal{X}_0 = A$, which shows that A is picked out.

To prove $V(\mathcal{H}) \leq d+1$, we need the following result from convex geometry; see e.g., https://en.wikipedia.org/wiki/Radon%27s_theorem.

By Talagrand's exponential bound (88),

$$\mathbb{P}\Big(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{H}} \geq x\Big) \leq \frac{D}{x}\left(\frac{Dx^2}{d+1}\right)^{d+1} \exp(-2x^2)$$

for all $n \geq 1$ *and* $x > 0$. *Taking* $x = \epsilon\sqrt{n}$ *yields*

$$\begin{aligned}\mathbb{P}\Big(\|\mathbb{P}_n - P\|_{\mathcal{H}} \geq \epsilon) &\leq \frac{D}{\epsilon\sqrt{n}}\left(\frac{D\epsilon^2 n}{d+1}\right)^{d+1} \exp(-2\epsilon^2 n) \\ &= \frac{D}{\epsilon\sqrt{n}} \exp\left((d+1)\log\left(\frac{D\epsilon^2 n}{d+1}\right) - 2\epsilon^2 n\right) \\ &\to 0 \qquad \text{as } n \to \infty,\end{aligned}$$

if $d/n \to 0$[71].

Lemma 7.24 (Radon's Lemma)**.** *Let* $\mathcal{X}_0 \subset \mathbb{R}^d$ *be a set of size* $d+2$. *Then there exist two disjoint subsets* $\mathcal{X}_1, \mathcal{X}_2$ *of* $\mathcal{X}_0$ *such that* $\operatorname{conv}(\mathcal{X}_1) \cap \operatorname{conv}(\mathcal{X}_2) \neq \emptyset$ *where* $\operatorname{conv}(\mathcal{X}_i)$ *denotes the convex hull of* $\mathcal{X}_i$.

Given Radon's lemma, the proof of $V(\mathcal{H}) \leq d+1$ is easy. We have to show that given any set $\mathcal{X}_0 \in \mathbb{R}^d$ of size $d+2$, there is a subset of $\mathcal{X}_0$ that cannot be picked out by the half-spaces. Using Radon's lemma with $\mathcal{X}_0$ yields two disjoint subsets $\mathcal{X}_1, \mathcal{X}_2$ of $\mathcal{X}_0$ such that $\operatorname{conv}(\mathcal{X}_1) \cap \operatorname{conv}(\mathcal{X}_2) \neq \emptyset$. We now claim that $\mathcal{X}_1$ cannot be picked out by using any half-space. Suppose that there is such a half-space H, i.e., $H \cap \mathcal{X}_0 = \mathcal{X}_1$. Note that if a half-space picks out a set of points, then every point in its convex hull is also picked out. Thus, $\operatorname{conv}(\mathcal{X}_1) \subset H$. However, as $\operatorname{conv}(\mathcal{X}_1) \cap \operatorname{conv}(\mathcal{X}_2) \neq \emptyset$, $H \cap \operatorname{conv}(\mathcal{X}_2) \neq \emptyset$ which implies that H also contains at least one point from $\mathcal{X}_2$, leading to a contradiction.

[71] Exercise (HW3): Show this.

8 Talagrand's concentration inequality for the suprema of the empirical process

The main goal of this chapter is to motivate and formally state (without proof) Talagrand's inequality for the suprema of the empirical process. We will also see a few applications of this result. If we have time, towards the end of the course, I will develop the tools necessary and prove the main result. To fully appreciate the strength of the main result, we start with a few important tail bounds for the sum of independent random variables. The following discussion extends and improves Hoeffding's inequality (Lemma 3.9).

In most of results in this chapter we only assume that the $\mathcal{X}$-valued random variables $X_1, \ldots, X_n$ are independent; they need not be identically distributed.

8.1 Preliminaries

Recall Hoeffding's inequality: Let $X_1, \ldots, X_n$ be independent and centered random variables such that $X_i \in [a_i, b_i]$ w.p.1 and let $S_n := \sum_{i=1}^n X_i$. Then, for any $t \geq 0$,

$$\mathbb{P}(S_n \geq t) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}, \quad \text{and} \quad \mathbb{P}(S_n \leq -t) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{89}$$

A crucial ingredient in the proof of the above result was Lemma 3.8 which stated that for a centered $X \in [a, b]$ w.p.1 we have $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 (b-a)^2/8}$, for $\lambda \geq 0$.

Note that if $b_i - a_i$ is much larger than the standard deviation σ_i of X_i then, although the tail probabilities prescribed by Hoeffding's inequality for S_n are of the normal type[72], they correspond to normal variables with the 'wrong' variance. The following result incorporates the standard deviation of the random variable and is inspired by the moment generating function of Poisson random variables[73].

Theorem 8.1. *Let X be a centered random variable such that $|X| \leq c$ a.s, for some $c < \infty$, and $\mathbb{E}[X^2] = \tau^2$. Then*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\tau^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right), \qquad \textit{for all } \lambda > 0. \tag{90}$$

As a consequence, if X_i, $1 \leq i \leq n$, are centered, independent and a.s. bounded by $c < \infty$ in absolute value, then setting

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2], \tag{91}$$

[72] Recall that if the X_i's are i.i.d. and centered with variance σ^2, by the CLT for fixed $t > 0$, $\lim_{n\to\infty} \mathbb{P}(S_n \geq t\sqrt{n}) = 1 - \Phi\left(\frac{t}{\sigma}\right) \leq \frac{\sigma}{\sqrt{2\pi}t} \exp\left(-\frac{t^2}{2\sigma^2}\right)$, where the last inequality uses a standard bound on the normal CDF.

[73] Recall that if X has Poisson distribution with parameter a (i.e., $\mathbb{E}X = \mathrm{Var}(X) = a$) then $\mathbb{E}[e^{\lambda(X-a)}] = e^{-a(\lambda+1)}\sum_{k=0}^\infty e^{\lambda k} a^k/k! = e^{a(e^\lambda - 1 - \lambda)}$.

and $S_n = \sum_{i=1}^n X_i$, *we have*

$$\mathbb{E}[e^{\lambda S_n}] \le \exp\left(\frac{n\sigma^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right), \qquad \text{for all } \lambda > 0, \tag{92}$$

and the same inequality holds for $-S_n$.

Proof. Since $\mathbb{E}(X) = 0$, expansion of the exponential gives

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!} \le \exp\Big(\sum_{i=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!}\Big).$$

Since $|\mathbb{E}X^k| \le c^{k-2}\tau^2$, for all $k \ge 2$, this exponent can be bounded by

$$\left|\sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!}\right| \le \lambda^2\tau^2 \sum_{k=2}^{\infty} \frac{(\lambda c)^{k-2}}{k!} = \frac{\tau^2}{c^2} \sum_{k=2}^{\infty} \frac{(\lambda c)^k}{k!} = \frac{\tau^2}{c^2}(e^{\lambda c} - 1 - \lambda c).$$

This gives inequality (90). Inequality (92) follows from (90) by using the independence of the X_i's. The above also applies to $Y_i = -X_i$ which yields the result for $-S_n$. □

It is standard to derive tail probability bounds for a random variable based on a bound for its moment generating function. We proceed to implement this idea and obtain four such bounds, three of them giving rise, respectively, to the *Bennett, Prokhorov* and *Bernstein* classical inequalities for sums of independent random variables and one where the bound on the tail probability function is inverted. It is convenient to introduce the following notation:

$$\begin{aligned} \phi(x) &= e^{-x} - 1 + x, \qquad \text{for } x \in \mathbb{R} \\ h_1(t) &= (1+t)\log(1+t) - t, \qquad \text{for } t \ge 0. \end{aligned}$$

Proposition 8.2. *Let* Z *be a random variable whose moment-generating function satisfies the bound*

$$\mathbb{E}(e^{\lambda Z}) \le \exp\big(\nu(e^{\lambda} - 1 - \lambda)\big), \qquad \lambda > 0, \tag{93}$$

for some $\nu > 0$. *Then, for all* $t \ge 0$,

$$\mathbb{P}(Z \ge t) \le e^{-\nu h_1(t/\nu)} \le \exp\left(-\frac{3t}{4}\log\left(1 + \frac{2t}{3\nu}\right)\right) \le \exp\left(-\frac{t^2}{2\nu + 2t/3}\right) \tag{94}$$

and

$$\mathbb{P}\Big(Z \ge \sqrt{2\nu x} + x/3\Big) \le e^{-x}, \qquad x \ge 0. \tag{95}$$

Proof. Observe that by Markov's inequality and the given bound $\mathbb{E}[e^{\lambda Z}]$, we obtain

$$\mathbb{P}(Z \ge t) = \inf_{\lambda > 0} \mathbb{P}(e^{\lambda Z} \ge e^{\lambda t}) \le \inf_{\lambda > 0} e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \le e^{\nu \inf_{\lambda > 0}\{\phi(-\lambda) - \lambda t/\nu\}}.$$

It can be checked that for $z > -1$ (think of $z = t/\nu$)

$$\inf_{\lambda \in \mathbb{R}}\{\phi(-\lambda) - \lambda z\} = z - (1+z)\log(1+z) = -h_1(z).$$

This proves the first inequality in (94). We can also show that (by checking the value of the corresponding functions at $t = 0$ and then comparing derivatives)

$$h_1(t) \geq \frac{3t}{4}\log\left(1+\frac{2t}{3}\right) \geq \frac{t^2}{2+2t/3}, \qquad \text{for } t > 0,$$

thus completing the proof of the three inequalities in (94).

To prove (95), we begin by observing that (by Taylor's theorem) $(1-\lambda/3)(e^\lambda - \lambda - 1) \leq \lambda^2/2, \lambda \geq 0$. Thus, if

$$\varphi(\lambda) := \frac{\nu\lambda^2}{2(1-\lambda/3)}, \qquad \lambda \in [0, 3),$$

then inequality (93) yields

$$\mathbb{P}(Z \geq t) \leq \inf_{0\leq\lambda<3} e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \leq \exp\left[\inf_{0\leq\lambda<3}(\varphi(\lambda)-\lambda t)\right] = \exp\left[-\sup_{0\leq\lambda<3}(\lambda t - \varphi(\lambda))\right] = e^{-\gamma(t)},$$

where we have used the fact that $\nu(e^\lambda - 1 - \lambda) \leq \varphi(\lambda)$ and $\gamma(s) := \sup_{\lambda\in[0,3)}(\lambda s - \varphi(\lambda))$, for $s > 0$. Then it can be shown[74] that $\gamma^{-1}(x) = \sqrt{2\nu x} + x/3$. Therefore, letting $t = \gamma^{-1}(x)$ (i.e., $x = \gamma(t)$) in the above display yields (95). □

Let $X_i, 1 \leq i \leq n$, be independent centered random variables a.s. bounded by $c < \infty$ in absolute value. Let $S_n := \sum_{i=1}^n X_i$ and define $Z := S_n/c$. Then,

$$\mathbb{E}[e^{\lambda Z}] = \prod_{i=1}^n \mathbb{E}[e^{(\lambda/c)X_i}] \leq \prod_{i=1}^n \exp\left(\frac{\mathbb{E}[X_i^2]}{c^2}(e^\lambda - 1 - \lambda)\right) = \exp\left(\frac{n\sigma^2}{c^2}(e^\lambda - 1 - \lambda)\right)$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2]$. Thus, Z satisfies the hypothesis of Proposition 8.2 with $\nu := n\sigma^2/c^2$. Therefore we have the following exponential inequalities, which go by the names of *Bennet's, Prokhorov's and Bernstein's*[75] (in that order).

Theorem 8.4. *Let $X_i, 1 \leq i \leq n$, be independent centered random variables a.s. bounded by $c < \infty$ in absolute value. Set $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]/n$ and $S_n := \sum_{i=1}^n X_i$. Then, for all $x \geq 0$,*

$$\mathbb{P}(S_n \geq t) \leq e^{-\left(\frac{n\sigma^2}{c^2}\right)h_1\left(\frac{tc}{n\sigma^2}\right)} \leq \exp\left(-\frac{3t}{4c}\log\left(1+\frac{2tc}{3n\sigma^2}\right)\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct/3}\right) \quad (96)$$

[74] Exercise (HW3): Complete this.

[75] It is natural to ask whether Theorem 8.4 extends to unbounded random variables. In fact, Bernstein's inequality does hold for random variables X_i with finite exponential moments, i.e., such that $\mathbb{E}[e^{\lambda|X_i|}] < \infty$, for some $\lambda > 0$, as shown below.

Lemma 8.3 (Bernstein's inequality). *Let $X_i, 1 \leq i \leq n$, be centered independent random variables such that, for all $k \geq 2$ and all $1 \leq i \leq n$,*

$$\mathbb{E}|X_i|^k \leq \frac{k!}{2}\sigma_i^2 c^{k-2},$$

and set $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$, $S_n := \sum_{i=1}^n X_i$. Then,

$$\mathbb{P}(S_n \geq t) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct}\right), \qquad \text{for } t \geq 0.$$

and

$$\mathbb{P}\Big(S_n \ge \sqrt{2n\sigma^2 x} + cx/3\Big) \le e^{-x}, \qquad x \ge 0.$$

Bennett's inequality is the sharpest, but Prokhorov's and Bernstein's inequalities are easier to interpret. Prokhorov's inequality exhibits two regimes for the tail probabilities of S_n: if $tc/(n\sigma^2)$ is small, then the logarithm is approximately $2tc/(3n\sigma^2)$, and the tail probability is only slightly larger than $e^{-t^2/(2n\sigma^2)}$ (which is Gaussian-like), whereas, if $tc/(n\sigma^2)$ is not small or moderate, then the exponent for the tail probability is of the order of $-[3t/(4c)]\log[2tc/(3n\sigma^2)]$ (which is 'Poisson'-like[76]). Bernstein's inequality keeps the Gaussian-like regime for small values of $tc/(n\sigma^2)$ but replaces the Poisson regime by the larger, hence less precise, exponential regime.

Example 8.5 (Deviation bound with fixed probability)**.** *Let us try to shed some light on the differences between Bernstein's inequality (i.e., the rightmost side of* (96)*) and Hoeffding's inequality (see* (89)*). We can first attempt to find the value of t which makes the bound on the rightmost side of* (96) *exactly equal to α, i.e., we want to solve the equation*

$$\exp\left(-\frac{t^2}{2(n\sigma^2 + ct/3)}\right) = \alpha.$$

This leads to the quadratic equation

$$t^2 - \frac{2tc}{3}\log\frac{1}{\alpha} - 2n\sigma^2\log\frac{1}{\alpha} = 0,$$

whose nonnegative solution is given by

$$t = \frac{c}{3}\log\frac{1}{\alpha} + \sqrt{\frac{c^2}{9}\left(\log\frac{1}{\alpha}\right)^2 + 2n\sigma^2\log\frac{1}{\alpha}} \;\;\le\;\; \sigma\sqrt{2n\log\frac{1}{\alpha}} + \frac{2c}{3}\log\frac{1}{\alpha}.$$

where in the last inequality we used the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$. Thus, Bernstein's inequality implies that $S_n \le \sigma\sqrt{2n\log\frac{1}{\alpha}} + \frac{2c}{3}\log\frac{1}{\alpha}$ with probability at least $1-\alpha$. Now if $X_1, \ldots, X_n$ are i.i.d. with mean zero, variance σ^2 and bounded in absolute value by c, then this yields

$$\bar{X}_n \le \frac{\sigma}{\sqrt{n}}\sqrt{2\log\frac{1}{\alpha}} + \frac{2c}{3n}\log\frac{1}{\alpha} \tag{97}$$

with probability (w.p.) at least $1-\alpha$; compare this the Hoeffding's bound which yields $\bar{X}_n \le c\sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ w.p. at least $1-\alpha$; see (11)*. Note that if $\bar{X}_n$ is normal, then $\bar{X}_n$ will be bounded by the first term in the right hand side of* (97) *w.p. at least $1-\alpha$. Therefore the above deviation bound agrees with the normal approximation bound except for the smaller order term (which if of order $1/n$; the leading term being of order $1/\sqrt{n}$).*

[76]Note that if X has Poisson distribution with parameter a (i.e., $\mathbb{E}X = \mathrm{Var}(X) = a$) then

$$\mathbb{P}(X - a \ge t) \le \exp\left[-\frac{3t}{4}\log\left(1 + \frac{2t}{3a}\right)\right], \qquad t \ge 0.$$

Example 8.6 (When X_i's are i.i.d. Bernoulli). *Suppose that X_i's are i.i.d. Bernoulli with probability of success $p \in (0,1)$. Then, using (97), we see that using the Bernstein's inequality yields that $\bar{X}_n \le \sqrt{\frac{p(1-p)}{n}}\sqrt{2\log\frac{1}{\alpha}} + \frac{2}{3n}\log\frac{1}{\alpha}$ holds w.p. at least $1-\alpha$; compare this with Hoeffding's inequality which yields $\bar{X}_n \le \sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ w.p. at least $1-\alpha$. Note that Bernstein's inequality is superior here if $p(1-p)$ is a fairly small. In particular, if $\mathrm{Var}(X_1) = \frac{1}{n}$ (i.e., $p \approx \frac{1}{n}$), then the two upper bounds reduce to $\frac{1}{n}\sqrt{2\log\frac{1}{\alpha}} + \frac{2}{3n}\log\frac{1}{\alpha}$ and $\sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ respectively, showing that Bernstein's inequality is so much better in this case.*

8.2 Talagrand's concentration inequality

Talagrand's concentration inequality for the supremum of the empirical process [Talagrand, 1996a] is one of the most useful results in modern empirical process theory, and also one of the deepest results in the theory. This inequality may be thought of as a Bennett, Prokhorov or Bernstein inequality uniform over an infinite collection of sums of independent random variables, i.e., for the supremum of the empirical process. As such, it constitutes an exponential inequality of the best possible kind. Below we state Bousquet's version of the upper half of Talagrand's inequality.

Theorem 8.7 (Talagrand's inequality, [Talagrand, 1996a, Bousquet, 2003]). *Let $X_i, i = 1,\ldots,n$, be independent $\mathcal{X}$-valued random variables. Let $\mathcal{F}$ be a countable family of measurable real-valued functions on $\mathcal{X}$ such that $\|f\|_\infty \le U < \infty$ and $\mathbb{E}[f(X_1)] = \ldots = \mathbb{E}[f(X_n)] = 0$, for all $f \in \mathcal{F}$. Let*

$$Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \qquad \text{or} \qquad Z = \sup_{f\in\mathcal{F}} \Big|\sum_{i=1}^n f(X_i)\Big|$$

and let the parameters σ^2 and ν_n be defined as

$$U^2 \ge \sigma^2 \ge \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_i)] \qquad \text{and} \qquad \nu_n := 2U\mathbb{E}[Z] + n\sigma^2.$$

Then[77]*, for all $t \ge 0$,*

$$\mathbb{P}(Z \ge \mathbb{E}Z + t) \le e^{-\left(\frac{\nu_n}{U^2}\right)h_1\left(\frac{tU}{\nu_n}\right)} \le \exp\left(-\frac{3t}{4U}\log\left(1+\frac{2tU}{3\nu_n}\right)\right) \le \exp\left(\frac{-t^2}{2\nu_n + 2tU/3}\right) \quad (98)$$

and

$$\mathbb{P}\Big(Z \ge \mathbb{E}Z + \sqrt{2\nu_n x} + Ux/3\Big) \le e^{-x}, \qquad x \ge 0. \quad (99)$$

[77] This is a consequence of the following: consider the class of functions $\tilde{\mathcal{F}} = \{f/U : f \in \mathcal{F}\}$ (thus any $\tilde{f} \in \tilde{\mathcal{F}}$ satisfies $\|\tilde{f}\|_\infty \le 1$). Let $\tilde{Z} := Z/U$, $\tilde{\sigma}^2 := \sigma^2/U^2$, and $\tilde{\nu}_n := \nu_n/U^2$. Then,

$$\log \mathbb{E}[e^{\lambda(\tilde{Z}-\mathbb{E}\tilde{Z})}] \le \tilde{\nu}_n(e^\lambda - 1 - \lambda), \qquad \lambda \ge 0.$$

Notice the similarity between (98) and the Bennet, Prokhorov and Bernstein inequalities in (96) in Theorem 8.4: in the case when $\mathcal{F} = \{f\}$, with $\|f\|_\infty \le c$, and $\mathbb{E}[f(X_i)] = 0$, U becomes c, and ν_n becomes $n\sigma^2$, and the right-hand side of Talagrand's inequality becomes exactly the Bennet, Prokhorov and Bernstein inequalities. Clearly, Talagrand's inequality is essentially the best possible exponential bound for the empirical process.

Whereas the Bousquet-Talagrand upper bound for the moment generating function of the supremum Z of an empirical process for $\lambda \ge 0$ is best possible, there exist quite good results for $\lambda < 0$, but these do not exactly reproduce the classical exponential bounds for sums of independent random variables when specified to a single function. Here is the strongest result available in this direction.

Theorem 8.8 ([Klein and Rio, 2005])**.** *Under the same hypothesis and notation as in Theorem 8.7, we have*

$$\log \mathbb{E}[e^{-\lambda(\tilde{Z}-\mathbb{E}\tilde{Z})}] \le \frac{\tilde{V}_n}{9}(e^{3\lambda} - 1 - 3\lambda), \qquad 0 \le \lambda < 1,$$

where $\tilde{V}_n = V_n/U^2$ and

$$V_n := 2U\mathbb{E}[Z] + \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E}[f^2(X_i)].$$

Then, for all $t \ge 0$,

$$\mathbb{P}(Z \le \mathbb{E}Z - t) \le e^{-\left(\frac{V_n}{9U^2}\right)h_1\left(\frac{3tU}{V_n}\right)} \le \exp\left(-\frac{t}{4U}\log\left(1 + \frac{2tU}{V_n}\right)\right) \le \exp\left(\frac{-t^2}{2V_n + 2tU}\right) \quad (100)$$

and

$$\mathbb{P}\Big(Z \le \mathbb{E}Z - \sqrt{2V_n x} - Ux\Big) \le e^{-x}, \qquad x \ge 0. \quad (101)$$

Remark 8.1. *In order to get concrete exponential inequalities from Theorems 8.7 and 8.8, we need to have good estimates of $\mathbb{E}Z$ and $\sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_i)]$. We have already seen many techniques to control $\mathbb{E}Z$. In particular, (85) gives such a bound.*

Example 8.9 (Dvoretzky-Kiefer-Wolfowitz)**.** *A first question we may ask is whether Talagrand's inequality recovers, up to constants, the DKW inequality. Let F be a distribution function in $\mathbb{R}^d$ and let $\mathbb{F}_n$ be the distribution function corresponding to n i.i.d. variables with distribution F. Let $Z := n\|\mathbb{F}_n - F\|_\infty$ We can take the envelope of the class $\mathcal{F} := \{\mathbf{1}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}^d\}$ to be 1 (i.e., $U = 1$), and $\sigma^2 = 1/4$. $\mathcal{F}$ is VC (with $V(\mathcal{F}) = d$) and inequality (85) gives*

$$\mathbb{E}[Z] = n\,\mathbb{E}\|\mathbb{F}_n - F\|_\infty \le c_1\sqrt{n},$$

where c_1 depends only on d. Here, $\nu_n \le 2c_1\sqrt{n} + n/4$. We have to upper-bound the probability

$$\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) = \mathbb{P}(Z \ge \sqrt{n}x) = \mathbb{P}(Z - \mathbb{E}Z \ge \sqrt{n}x - \mathbb{E}Z).$$

Note that for $x > 2\sqrt{n}$, this probability is zero (as $Z \le 2n$). For $x > 2c_1$, $t := \sqrt{n}x - \mathbb{E}Z \ge \sqrt{n}(x - c_1) > 0$, and thus we can apply the last inequality in (98). Hence, for $2\sqrt{n} \ge x > 2c_1$,

$$\begin{aligned}\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) &\le \exp\left(-\frac{(\sqrt{n}x - \mathbb{E}Z)^2}{2(2c_1\sqrt{n} + n/4) + 2(\sqrt{n}x - \mathbb{E}Z)/3}\right)\\ &\le \exp\left(-\frac{n(x-c_1)^2}{c_3 n}\right) \le \exp\left(-\frac{x^2}{4c_3}\right),\end{aligned}$$

where we have used (i) for $2\sqrt{n} \ge x$ the denominator in the exponential term is upper bounded by $2(2c_1\sqrt{n}+n/4)+4n/3$ which is in turn upper bounded by $c_3 n$ (for some $c_3 > 0$); (ii) for $x > 2c_1$, $(x-c_1)^2 > x^2/4$ (as $x - c_1 \ge x - x/2 = x/2$). Thus, for some constants $c_2, c_3 > 0$ that depend only on d, we can show that for all $x > 0$,

$$\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) \le c_2 e^{-x^2/(4c_3)};$$

note that c_2 appears above as we can just upper bound the LHS by some c_2 when $x \in (0, 2c_1]$.

Example 8.10 (Data-driven inequalities)**.** *In many statistical applications, it is of importance to have data-dependent "confidence sets" for the random quantity $\|\mathbb{P}_n - P\|_{\mathcal{F}}$. This quantity is a natural measure of the accuracy of the approximation of an unknown distribution by the empirical distribution $\mathbb{P}_n$. However, $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ itself depends on the unknown distribution P and is not directly available.*

To obtain such data dependent bounds on $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ we have to replace the unknown quantities $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$, σ^2 and U by suitable estimates or bounds. Suppose for the sake of simplicity, σ^2 and U are known, and the only problem is to estimate or bound the expectation $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$. We have discussed so far how to bound the expectation $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$. However, such bounds typically depend on other unknown constants and may not be sharp. Talagrand's inequalities (99) and (101), and symmetrization allow us to replace $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$ by a completely data-based surrogate. In the following we give such a (finite-sample) high-probability upper bound on $\|\mathbb{P}_n - P\|_{\mathcal{F}}$; see [Giné and Nickl, 2016, Section 3.4.2] for more on this topic.

Theorem 8.11. *Let $\mathcal{F}$ be a countable collection of real-valued measurable functions on $\mathcal{X}$ with absolute values bounded by 1/2. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$ with a common probability law P. Let $\varepsilon_1, \ldots, e_n$ be i.i.d. Rademacher random variables independent from the sequence $\{X_i\}$ and let $\sigma^2 \ge \sup_{f\in\mathcal{F}} Pf^2$. Then, for all n and $x \ge 0$,*

$$\mathbb{P}\left(\|\mathbb{P}_n - P\|_{\mathcal{F}} \ge 3\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}} + 4\sqrt{\frac{2\sigma^2 x}{n}} + \frac{70}{3n}x\right) \le 2e^{-x}.$$

Proof. Set $Z := \|\sum_{i=1}^n (f(X_i) - Pf)\|_{\mathcal{F}}$ and set $\tilde{Z} := \|\sum_{i=1}^n \epsilon_i f(X_i)\|_{\mathcal{F}}$. Note that $\tilde{Z}$ is also the supremum of an empirical process: the variables are $\tilde{X}_i = (\varepsilon_i, X_i)$, defined on $\{-1,1\} \times \mathcal{X}$, and the functions are $\tilde{f}(\epsilon, x) := \epsilon f(x)$, for $f \in \mathcal{F}$. Thus, Talagrand's inequalities apply to both Z and $\tilde{Z}$. Then, using the fact

$$\sqrt{2x(n\sigma^2 + 2\mathbb{E}\tilde{Z})} \leq \sqrt{2xn\sigma^2} + 2\sqrt{x\mathbb{E}\tilde{Z}} \leq \sqrt{2xn\sigma^2} + \frac{1}{\delta}x + \delta\mathbb{E}\tilde{Z},$$

for any $\delta > 0$, the Klein-Rio version of Talagrand's lower-tail inequality gives

$$e^{-x} \geq \mathbb{P}\left(\tilde{Z} \leq \mathbb{E}\tilde{Z} - \sqrt{2x(n\sigma^2 + 2\mathbb{E}\tilde{Z})} - x\right) \geq \mathbb{P}\left(\tilde{Z} \leq (1-\delta)\mathbb{E}\tilde{Z} - \sqrt{2xn\sigma^2} - \frac{1+\delta}{\delta}x\right).$$

Similarly, using (99),

$$\mathbb{P}\left(Z \geq (1+\delta)\mathbb{E}Z + \sqrt{2xn\sigma^2} + \frac{3+\delta}{3\delta}x\right) \leq e^{-x}.$$

Recall also that $\mathbb{E}[Z] \leq 2\mathbb{E}[\tilde{Z}]$. Then, we have on the intersection of the complement of the events in the last two inequalities, for $\delta = 1/5$ (say),

$$\begin{aligned} Z &< \frac{6}{5}\mathbb{E}[Z] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \leq \frac{12}{5}\mathbb{E}[\tilde{Z}] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \\ &< \frac{12}{5}\left[\frac{5}{4}\tilde{Z} + \frac{5}{4}\sqrt{2xn\sigma^2} + \frac{15}{2}x\right] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \\ &= 3\tilde{Z} + 4\sqrt{2xn\sigma^2} + \frac{70}{3}x; \end{aligned}$$

i.e., this inequality holds with probability $1 - 2e^{-x}$. $\square$

Note that different values of δ produce different coefficients in the above theorem.

8.3 Empirical risk minimization and concentration inequalities

Let $X, X_1, \ldots, X_n, \ldots$ be i.i.d. random variables defined on a probability space and taking values in a measurable space $\mathcal{X}$ with common distribution P. In this section we highlight the usefulness of concentration inequalities, especially Talagrand's inequality, in empirical risk minimization (ERM); see [Koltchinskii, 2011] for a thorough study of this topic.

Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. In what follows, the values of a function $f \in \mathcal{F}$ will be interpreted as "losses" associated with certain "actions" (e.g., $\mathcal{F} = \{f(x) \equiv f(z,y) = (y - \beta^\top z)^2 : \beta \in \mathbb{R}^d\}$ and $X = (Z, Y) \sim P$).

We will be interested in the problem of risk minimization:

$$\min_{f \in \mathcal{F}} Pf \tag{102}$$

in the cases when the distribution P is unknown and has to be estimated based on the data $X_1, \ldots, X_n$. Since the empirical measure $\mathbb{P}_n$ is a natural estimator of P, the true risk can be estimated by the corresponding empirical risk, and the risk minimization problem has to be replaced by the *empirical risk minimization* (ERM):

$$\min_{f \in \mathcal{F}} \mathbb{P}_n f. \tag{103}$$

As is probably clear by now, many important methods of statistical estimation such as maximum likelihood and more general M-estimation are versions of ERM.

Definition 8.12. *The excess risk of $f \in \mathcal{F}$ is defined as*

$$\mathcal{E}(f) \equiv \mathcal{E}_P(f) := Pf - \inf_{h \in \mathcal{F}} Ph.$$

Recall that we have already seen an important application of ERM in the problem of classification in Example 7.10. Here is another important application.

Example 8.13 (Regression)**.** *Suppose that we observe $X_1 \equiv (Z_1, Y_1), \ldots, X_n \equiv (Z_n, Y_n)$ i.i.d. $X \equiv (Z, Y) \sim P$ on $\mathcal{X} \equiv \mathcal{Z} \times T$, $T \subset \mathbb{R}$, and the goal is to study the relationship between Y and Z. We study regression with quadratic loss $\ell(y, u) := (y - u)^2$ given a class of of measurable functions $\mathcal{G}$ from $\mathcal{Z}$ to T; the distribution of Z will be denoted by Π. This problem can be thought of as a special case of ERM with*

$$\mathcal{F} := \{(\ell \bullet g)(z, y) \equiv (y - g(z))^2 : g \in \mathcal{G}\}.$$

Suppose that the true regression function is $g_*(z) := \mathbb{E}[Y|Z = z]$, for $z \in \mathcal{Z}$. In this case, the excess risk of $f(z, y) = (y - g(z))^2 \in \mathcal{F}$ (for some $g \in \mathcal{G}$) is given by[78]

$$\mathcal{E}_P(f) = \mathcal{E}_P(\ell \bullet g) = \|g - g_*\|^2_{L_2(\Pi)} - \inf_{h \in \mathcal{G}} \|h - g_*\|^2_{L_2(\Pi)}. \tag{104}$$

If $\mathcal{G}$ is such that $g_* \in \mathcal{G}$ then $\mathcal{E}_P(\ell \bullet g) = \|g - g_*\|^2_{L_2(\Pi)}$, for all $g \in \mathcal{G}$.

Let

$$\hat{f} \equiv \hat{f}_n \in \arg\min_{f \in \mathcal{F}} \mathbb{P}_n f$$

be a solution of the ERM problem (103). The function $\hat{f}_n$ is used as an approximation of the solution of the true risk minimization problem (102) and its excess risk $\mathcal{E}_P(\hat{f}_n)$ is a natural measure of accuracy of this approximation.

It is worth pointing out that a crucial difference between ERM and classical M-estimation, as discussed in Sections 5 and 6, is that in the analysis of ERM we do not (usually) assume that the data generating distribution P belongs to the class of models considered (e.g., $\inf_{h \in \mathcal{F}} Ph$ need not be 0). Moreover, in M-estimation, typically the focus is on recovering a parameter of interest in the model (which is expressed as the population M-estimator) whereas in ERM the focus is mainly on deriving optimal (upper and lower) bounds for the excess risk $\mathcal{E}_P(\hat{f}_n)$.

It is of interest to find tight upper bounds on the excess risk[79] of $\hat{f}$ that hold with a high probability. Such bounds usually depend on certain "geometric" properties of the

[78] Exercise (HW3): Show this.

[79] Note that we have studied upper bounds on the excess risk in the problem of classification in Example 7.10.

function class $\mathcal{F}$ and on various measures of its "complexity" that determine the accuracy of approximation of the true risk Pf by the empirical risk $\mathbb{P}_n f$ in a neighborhood of a proper size of the minimal set of the true risk.

In the following we describe a rather general approach to derivation of such bounds in an abstract framework of ERM. We start with some definitions.

Definition 8.14. *The δ-minimal set of the risk is defined as*

$$\mathcal{F}(\delta) := \{f \in \mathcal{F} : \mathcal{E}_P(f) \le \delta\}.$$

The L_2-diameter of the δ-minimal set is denoted by

$$D(\delta) \equiv D_P(\mathcal{F};\delta) := \sup_{f_1,f_2 \in \mathcal{F}(\delta)} \{P[(f_1 - f_2)^2]\}^{1/2}.$$

Suppose, for simplicity, that the infimum of the risk Pf is attained at $\bar{f} \in \mathcal{F}$ (the argument can be easily modified if the infimum is not attained in the class). Denote

$$\hat{\delta} := \mathcal{E}_P(\hat{f}).$$

Then $\hat{f}, \bar{f} \in \mathcal{F}(\hat{\delta})$ and $\mathbb{P}_n \hat{f} \le \mathbb{P}_n \bar{f}$. Therefore,

$$\begin{aligned} \hat{\delta} &= \mathcal{E}_P(\hat{f}) = P(\hat{f} - \bar{f}) \le \mathbb{P}_n(\hat{f} - \bar{f}) + (P - \mathbb{P}_n)(\hat{f} - \bar{f}) \\ &\le \sup_{f_1,f_2 \in \mathcal{F}(\hat{\delta})} |(\mathbb{P}_n - P)(f_1 - f_2)| \qquad (105) \\ &\le \sup_{f_1,f_2 \in \mathcal{F}} |(\mathbb{P}_n - P)(f_1 - f_2)|. \end{aligned}$$

Previously, we had used the last inequality to upper bound the excess risk in classification; see Example 7.10. In this section we will use the implicit characterization of $\hat{\delta}$ in (105) to improve our upper bound. This naturally leads us to the study of the following (local) measure of empirical approximation:

$$\phi_n(\delta) \equiv \phi_n(\mathcal{F};\delta) := \mathbb{E}\left[\sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)|\right]. \qquad (106)$$

Idea: Imagine there exists a nonrandom upper bound

$$U_n(\delta) \ge \sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \qquad (107)$$

that holds *uniformly* in δ with a high probability. Then, with the same probability, the excess risk $\hat{\delta} = \mathcal{E}_P(\hat{f})$ will be bounded[80] by the largest solution of the inequality

$$\delta \le U_n(\delta). \qquad (108)$$

[80] As $\hat{\delta} \le \sup_{f_1,f_2 \in \mathcal{F}(\hat{\delta})} |(\mathbb{P}_n - P)(f_1 - f_2)| \le U_n(\hat{\delta})$, $\hat{\delta}$ satisfies inequality (108).

By solving the above inequality one can obtain $\delta_n(\mathcal{F})$ (which satisfies (108)) such that $\mathbb{P}(\mathcal{E}_P(\hat{f}_n) > \delta_n(\mathcal{F}))$ is small[81]. Thus, constructing an upper bound on the excess risk essentially reduces to solving a fixed point inequality of the type $\delta \leq U_n(\delta)$.

Let us describe in more detail what we mean by the above intuition. There are many different ways to construct upper bounds on the sup-norm of empirical processes. A very general approach is based on Talagrand's concentration inequalities. For example, if the functions in $\mathcal{F}$ take values in the interval $[0, 1]$, then[82] by (99) we have, for $t > 0$,[83]

$$\mathbb{P}\left(\sup_{f_1, f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \geq \phi_n(\delta) + \frac{1}{\sqrt{n}}\sqrt{2t\left(2\phi_n(\delta) + D^2(\delta)\right)} + \frac{t}{3n}\right) \leq e^{-t}. \quad (109)$$

Then, using the facts: (i) $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and (ii) $2\sqrt{ab} \leq a/K + Kb$, for any $a, b, K > 0$, we have

$$\sqrt{2t\left(D^2(\delta) + 2\phi_n(\delta)\right)} \leq \sqrt{2tD^2(\delta)} + 2\sqrt{t\phi_n(\delta)} \leq D(\delta)\sqrt{2t} + \frac{t}{\sqrt{n}} + \sqrt{n}\phi_n(\delta).$$

Thus, from (109), for all $t > 0$, we have[84]

$$\mathbb{P}\left(\sup_{f_1, f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \geq \bar{U}_n(\delta; t)\right) \leq e^{-t} \quad (110)$$

where

$$\bar{U}_n(\delta; t) := 2\left(\phi_n(\delta) + D(\delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\right). \quad (111)$$

This observation provides a way to construct a function $U_n(\delta)$ such that (107) holds with a high probability "uniformly" in δ — by first defining such a function at a discrete set of values of δ and then extending it to all values by monotonicity. We will elaborate on this shortly. Then, by solving the inequality (108) one can construct a bound on $\mathcal{E}_P(\hat{f}_n)$, which holds with "high probability" and which is often of correct order of magnitude.

8.3.1 A formal result on excess risk in ERM

Let us now try to state a formal result in this direction. To simplify notation, assume that the functions in $\mathcal{F}$ take values in $[0, 1]$. Let $\{\delta_j\}_{j \geq 0}$ be a decreasing sequence of positive numbers with $\delta_0 = 1$ and let $\{t_j\}_{j \geq 0}$ be a sequence of positive numbers. Define $U_n : (0, \infty) \to \mathbb{R}$, via (111), as

$$U_n(\delta) := \bar{U}_n(\delta_j; t_j), \qquad \text{for} \quad \delta \in (\delta_{j+1}, \delta_j], \quad (112)$$

[81] We will formalize this later.

[82] This assumption just simplifies a few mathematical expressions; there is nothing sacred about the interval $[0, 1]$, we could have done it for any constant compact interval.

[83] According to the notation of (99), we can take $\sigma^2 = D^2(\delta)$, and then $\nu_n = 2n\phi_n(\mathcal{F}; \delta) + nD^2(\delta)$.

[84] This form of the concentration inequality is usually called Bousquet's version of Talagrand's inequality.

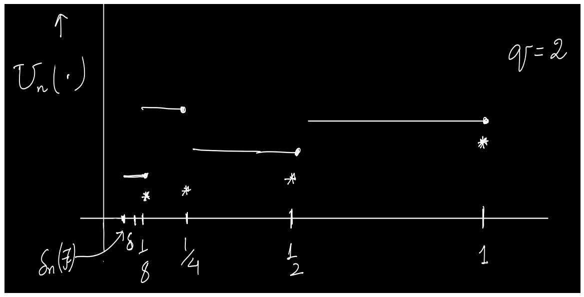


Figure 2: Plot of the piecewise constant function $U_n(\delta)$, for $\delta \geq \delta_n(\mathcal{F})$, along with the value of $\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)}$, for $j = 0, 1, \ldots$, denoted by the $\star$'s.

and $U_n(\delta) := U_n(1)$ for $\delta > 1$. Denote

$$\delta_n(\mathcal{F}) := \sup\{\delta \in (0, 1] : \delta \leq U_n(\delta)\}. \tag{113}$$

It is easy to check that $\delta_n(\mathcal{F}) \leq U_n(\delta_n(\mathcal{F}))$. Obviously, the definitions of U_n and $\delta_n(\mathcal{F})$ depend on the choice of $\{\delta_j\}_{j\geq 0}$ and $\{t_j\}_{j\geq 0}$ (we will choose specific values of these quantities later on). We start with the following simple inequality that provides a distribution dependent upper bound on the excess risk $\mathcal{E}_P(\hat{f}_n)$.

Theorem 8.15. *For all $\delta \geq \delta_n(\mathcal{F})$,*

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \delta\right) \leq \sum_{j:\delta_j \geq \delta} e^{-t_j}. \tag{114}$$

Proof. It is enough to prove the result for any $\delta > \delta_n(\mathcal{F})$; then the right continuity of the distribution function of $\mathcal{E}_P(\hat{f}_n)$ would lead to the bound (114) for $\delta = \delta_n(\mathcal{F})$.

So, fix $\delta > \delta_n(\mathcal{F})$. Letting $\mathcal{F}'(\delta) := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}(\delta)\}$, we know that

$$\mathcal{E}_P(\hat{f}) = \hat{\delta} \leq \sup_{f \in \mathcal{F}'(\hat{\delta})} |(\mathbb{P}_n - P)(f)| \equiv \|\mathbb{P}_n - P\|_{\mathcal{F}'(\hat{\delta})}. \tag{115}$$

Denote

$$E_{n,j} := \left\{\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)} \leq U_n(\delta_j)\right\}.$$

It follows from Bousquet's version of Talagrand's inequality (see (110)) that $\mathbb{P}(E_{n,j}) \geq 1 - e^{-t_j}$. Let

$$E_n := \cap_{j:\delta_j \geq \delta} E_{n,j}.$$

Then

$$\mathbb{P}(E_n) = 1 - \mathbb{P}(E_n^c) \geq 1 - \sum_{j:\delta_j \geq \delta} e^{-t_j}. \tag{116}$$

On the event E_n, for all $\sigma \geq \delta$, we have

$$\|\mathbb{P}_n - P\|_{\mathcal{F}'(\sigma)} \leq U_n(\sigma). \tag{117}$$

The above holds as: (i) $U_n(\cdot)$ is a piecewise constant function (with possible jumps only at δ_j's), (ii) the function $\sigma \mapsto \|\mathbb{P}_n - P\|_{\mathcal{F}'(\sigma)}$ is monotonically nondecreasing, and (iii) $\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)} \le U_n(\delta_j)$ on E_n, for j such that $\delta \ge \delta_j$; see Figure 8.3.1.

Claim: $\{\hat{\delta} \ge \delta\} \subset E_n^c$. We prove the claim using the method of contradiction. Thus, suppose that the above claim does not hold. Then, the event $\{\hat{\delta} \ge \delta\} \cap E_n$ is non-empty. On the event $\{\hat{\delta} \ge \delta\} \cap E_n$ we have

$$\hat{\delta} \le \|\mathbb{P}_n - P\|_{\mathcal{F}'(\hat{\delta})} \le U_n(\hat{\delta}), \tag{118}$$

where the first inequality follows from (115) and the second inequality holds via (117). This, in particular, implies that

$$\delta \le \hat{\delta} \le \delta_n(\mathcal{F}),$$

where the last inequality follows from (118) and the maximality of $\delta_n(\mathcal{F})$ via (113). However the above display contradicts the assumption that $\delta > \delta_n(\mathcal{F})$. Therefore, we must have $\{\hat{\delta} \ge \delta\} \subset E_n^c$.

The claim now implies that $\mathbb{P}(\mathcal{E}_P(\hat{f}_n) \ge \delta) = \mathbb{P}(\hat{\delta} \ge \delta) \le \mathbb{P}(E_n^c) \le \sum_{j:\delta_j \ge \delta} e^{-t_j}$, via (116), thereby completing the proof. $\square$

Although Theorem 8.15 yields a high probability bound on the excess risk of $\hat{f}_n$ (i.e., $\mathcal{E}_P(\hat{f}_n)$), we still need to upper bound $\delta_n(\mathcal{F})$ for the result to be useful. We address this next. We start with some notation. Given any $\psi : (0, \infty) \to \mathbb{R}$, denote by

$$\psi^\dagger(\sigma) := \sup_{s \ge \sigma} \frac{\psi(s)}{s}. \tag{119}$$

Note that $\psi^\dagger$ is a nonincreasing function[85].

The study of $\psi^\dagger$ is naturally motivated by the study of the function $\frac{U_n(\delta)}{\delta}$ and when it crosses the value 1; cf. (113). As $\frac{U_n(\delta)}{\delta}$ may have multiple crossings of 1, we "regularize" $\frac{U_n(\delta)}{\delta}$ by studying $V_n^t(\delta)$ defined below (which can be thought of as a well-behaved monotone version of $U_n^\dagger$). For $q > 1$ an $t > 0$, denote

$$V_n^t(\sigma) := 2q\left[\phi_n^\dagger(\sigma) + \sqrt{(D^2)^\dagger(\sigma)}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\right], \qquad \text{for } \sigma > 0. \tag{120}$$

Note that V_n^t is a strictly decreasing of σ in $(0, \infty)$. Let

$$\sigma_n^t \equiv \sigma_n^t(\mathcal{F}) := \inf\{\sigma > 0 : V_n^t(\sigma) \le 1\}. \tag{121}$$

We will show next that $\sigma_n^t \ge \delta_n(\mathcal{F})$ (for a special choice of $\{\delta_j\}_{j\ge 0}$ and $\{t_j\}_{j\ge 0}$) and thus, by (8.15) and some algebraic simplification, we will obtain the following result. Given

[85] Take $\sigma_1 < \sigma_2$. Then

$$\psi^\dagger(\sigma_1) = \sup_{s \ge \sigma_1} \frac{\psi(s)}{s} \ge \sup_{s \ge \sigma_2} \frac{\psi(s)}{s} = \psi^\dagger(\sigma_2).$$

a concrete application, our goal would be to find upper bounds on σ_n^t; see Section 8.3.2 where we illustrate this technique for finding a high probability bound on the excess risk in bounded regression.

Theorem 8.16 (High probability bound on the excess risk of the ERM)**.** *For all $t > 0$,*

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \sigma_n^t\right) \leq C_q e^{-t}. \tag{122}$$

where $C_q := \frac{q}{q-1} \vee e$.

Proof. Fix $t > 0$ and let $\sigma > \sigma_n^t$. We will show that $\mathbb{P}\big(\mathcal{E}_P(\hat{f}_n) > \sigma\big) \leq C_q e^{-t}$. Then, by taking a limit as $\sigma \downarrow \sigma_n^t$, we obtain (122).

Define, for $j \geq 0$,

$$\delta_j := q^{-j} \qquad \text{and} \qquad t_j := t\frac{\delta_j}{\sigma}.$$

Recall the definitions of $U_n(\delta)$ and $\delta_n(\mathcal{F})$ (in (112) and (113)) using the above choice of the sequences $\{\delta_j\}_{j\geq 0}$ and $\{t_j\}_{j\geq 0}$. Then, for all $\delta \geq \sigma$, using (112),[86]

$$\begin{aligned}
\frac{U_n(\delta)}{\delta} &= 2\left(\frac{\phi_n(\delta_j)}{\delta} + \frac{D(\delta_j)}{\sqrt{\delta}}\sqrt{\frac{t\delta_j}{\delta\sigma n}} + \frac{t\delta_j}{\delta\sigma n}\right) && \text{if } \delta \in (\delta_{j+1}, \delta_j] \\
&\leq 2q\left(\frac{\phi_n(\delta_j)}{\delta_j} + \frac{D(\delta_j)}{\sqrt{\delta_j}}\sqrt{\frac{t\delta_j}{\delta_j\sigma n}} + \frac{t\delta_j}{\delta_j\sigma n}\right) && \text{as } \delta > \delta_{j+1} = \frac{\delta_j}{q} \Rightarrow \frac{1}{\delta} < \frac{q}{\delta_j} \\
&\leq 2q\left(\sup_{s\geq\sigma}\frac{\phi_n(s)}{s} + \sqrt{\frac{t}{\sigma n}}\sup_{s\geq\sigma}\frac{D(\delta)}{\sqrt{\delta}} + \frac{t}{\sigma n}\right) && \text{as } \delta_j \geq \delta \geq \sigma \\
&= 2q\left(\phi_n^\dagger(\sigma) + \sqrt{(D^2)^\dagger(\sigma)}\sqrt{\frac{t}{\sigma n}} + \frac{t}{\sigma n}\right) = V_n^t(\sigma).
\end{aligned}$$

Since $\sigma > \sigma_n^t$ and the function V_n^t is strictly decreasing, we have $V_n^t(\sigma) < V_n^t(\sigma_n^t) \leq 1$, and hence, for all $\delta > \sigma$,

$$\frac{U_n(\delta)}{\delta} \leq V_n^t(\sigma) < 1.$$

Therefore, $\delta > \delta_n(\mathcal{F}) := \sup\{s > 0 : 1 \leq \frac{U_n(s)}{s}\}$, and thus, $\sigma \geq \delta_n(\mathcal{F})$. Now, from Theorem 8.15 it follows that

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \sigma\right) \leq \sum_{j:\delta_j\geq\sigma} e^{-t_j} \leq C_q e^{-t}$$

where the last step follows from some algebra[87]. □

[86]For $\delta > \delta_0 \equiv 1$, the following sequence of displays also holds with $j = 0$.

[87]Exercise (HW3): Show this. Hint: we can write

$$\sum_{j:\delta_j\geq\sigma} e^{-t_j} = \sum_{j:\delta_j\geq\sigma} e^{-t\delta_j/\sigma} \leq \sum_{j\geq 0} e^{-tq^j} = \cdots \leq \frac{q}{q-1}e^{-t}, \qquad \text{for } t \geq 1.$$

8.3.2 Excess risk in bounded regression

Recall the regression setting in Example 8.13. Given a function $g : \mathcal{Z} \to T$, the quantity $(\ell \bullet g)(z, y) := \ell(y, g(z))$ is interpreted as the loss suffered when $g(z)$ is used to predict y. The problem of optimal prediction can be viewed as a *risk minimization*:

$$\mathbb{E}[\ell(Y, g(Z))] =: P(\ell \bullet g)$$

over $g : \mathcal{Z} \to T$. We start with the regression problem with *bounded response* and with quadratic loss. To be specific, assume that Y takes values in $T = [0, 1]$ and $\ell(y, u) := (y-u)^2$. Suppose that we are given a class of measurable real-valued functions $\mathcal{G}$ on $\mathcal{Z}$. We denote by $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$. Suppose that the true regression function is $g_*(z) := \mathbb{E}[Y|Z = z]$, for $z \in \mathcal{Z}$, which is not assumed to be in $\mathcal{G}$. Recall that the *excess risk* $\mathcal{E}_P(\ell \bullet g)$ in this problem is given by (104).

In order to apply Theorem 8.16 to find a high probability bound on the excess risk of the ERM $\hat{f} \equiv \ell \bullet \hat{g}$ (see (103)) in this problem, which is determined by σ_n^t via (121), we have to find upper bounds for $V_n^t(\cdot)$ (which in turn depends on the functions $\phi_n^\dagger$ and $\sqrt{(D^2)^\dagger}$).

As a first step we relate the excess risk of any $f \equiv \ell \bullet g \in \mathcal{F}$ to $g \in \mathcal{G}$. The following lemma provides an easy way to bound the excess risk of f from below in the case of a *convex class* $\mathcal{G}$, an assumption we make in the sequel.

Lemma 8.17. *If $\mathcal{G}$ is a convex class of functions, then*

$$2\mathcal{E}_P(\ell \bullet g) \geq \|g - \bar{g}\|^2_{L_2(\Pi)}$$

where $\bar{g} := \operatorname{argmin}_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)}$ is assumed to exist.

Below we make some observations that will be crucial to find σ_n^t.

1. It follows from Lemma 8.17 that

$$\mathcal{F}(\delta) = \{f \in \mathcal{F} : \mathcal{E}_P(f) \leq \delta\} \subset \{\ell \bullet g : g \in \mathcal{G},\ \|g - \bar{g}\|^2_{L_2(\Pi)} \leq 2\delta\}. \tag{123}$$

2. For any two functions $g_1, g_2 \in \mathcal{G}$ and all $z \in \mathcal{Z}, y \in [0, 1]$, we have

$$\begin{aligned} |(\ell \bullet g_1)(z, y) - (\ell \bullet g_2)(z, y)| &= \left|(y - g_1(z))^2 - (y - g_2(z))^2\right| \\ &= |g_1(z) - g_2(z)|\,|2y - g_1(z) - g_2(z)| \ \leq\ 2\,|g_1(z) - g_2(z)|\,, \end{aligned}$$

   which implies

$$P\left[(\ell \bullet g_1 - \ell \bullet g_2)^2\right] \leq 4\|g_1 - g_2\|^2_{L_2(\Pi)}.$$

   Recalling that $D(\delta) := \sup_{f_1, f_2 \in \mathcal{F}(\delta)} \{P[(f_1 - f_2)^2]\}^{1/2}$, we have

$$\begin{aligned} D(\delta) &\leq 2 \sup\left\{\|g_1 - g_2\|_{L_2(\Pi)} : g_k \in \mathcal{G}, \|g_k - \bar{g}\|^2_{L_2(\Pi)} \leq 2\delta \text{ for } k = 1, 2\right\} \\ &\leq 2(2\sqrt{2}\delta) \end{aligned} \tag{124}$$

where the last step follows from the triangle inequality: $\|g_1 - g_2\|_{L_2(\Pi)} \le \|g_1 - \bar{g}\|_{L_2(\Pi)} + \|g_2 - \bar{g}\|_{L_2(\Pi)}$. Hence, by (124),

$$\sqrt{(D^2)^\dagger(\sigma)} = \sqrt{\sup_{\delta \ge \sigma} \frac{D^2(\delta)}{\delta}} \le 4\sqrt{2}.$$

3. By symmetrization inequality (recall that we use $\epsilon_1, \ldots, \epsilon_n$ to be i.i.d. Rademacher variables independent of the observed data), and letting $\mathcal{F}'(\delta) := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}(\delta)\}$, and using (123),

$$\begin{aligned} \phi_n(\delta) = \mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta)} &\le 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}'(\delta)} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i f(X_i)\Big|\right] \\ &\le 2\,\mathbb{E}\left[\sup_{g_k \in \mathcal{G}: \|g_k - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\,(\ell \bullet g_1 - \ell \bullet g_2)(X_i)\Big|\right] \\ &\le 4\,\mathbb{E}\left[\sup_{g \in \mathcal{G}: \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\,(\ell \bullet g - \ell \bullet \bar{g})(X_i)\Big|\right]. \end{aligned}$$

Since $\ell(y, \cdot)$ is Lipschitz with constant 2 on the interval $[0,1]$ one can use the *contraction inequality*[88] to get

$$\phi_n(\delta) \le 8\,\mathbb{E}\left[\sup_{g \in \mathcal{G}: \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i (g - \bar{g})(Z_i)\Big|\right] := \psi_n(\delta).$$

As a result, we get (recall (119))

$$\phi_n^\dagger(\sigma) \le \psi_n^\dagger(\sigma).$$

The following result is now a corollary of Theorem 8.16.

Theorem 8.18. *Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ into $[0,1]$ and let $\hat{g}_n$ denotes the LSE of the regression function, i.e.,*

$$\hat{g}_n := \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \{Y_i - g(X_i)\}^2.$$

Then, there exist constants $K > 0$ such that for all $t > 0$,

$$\mathbb{P}\left\{\|\hat{g}_n - g_*\|^2_{L_2(\Pi)} \ge \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + \left(\psi_n^\sharp(\frac{1}{4q}) + K\frac{t}{n}\right)\right\} \le C_q e^{-t}, \qquad (125)$$

[88] Ledoux-Talagrand contraction inequality (Theorem 4.12 of [Ledoux and Talagrand, 1991]): If $\varphi_i : \mathbb{R} \to \mathbb{R}$ satisfies $|\varphi_i(a) - \varphi_i(b)| \le L|a - b|$ for all $a, b \in \mathbb{R}$, then

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_i(h(x_i))\right] \le L\,\mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i)\right].$$

In the above application we take $\varphi_i(u) = (Y_i - u)^2$ for $u \in [0,1]$.

where for any $\psi : (0, \infty) \to \mathbb{R}$, $\psi^\sharp$ *is defined as*[89]

$$\psi^\sharp(\varepsilon) := \inf\left\{\sigma > 0 : \psi^\dagger(\sigma) \le \varepsilon\right\}. \tag{126}$$

Proof. Note that in this case, by (104), $\mathcal{E}_P(\hat{g}_n) = \|\hat{g}_n - g_*\|^2_{L_2(\Pi)} - \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)}$. To use Theorem 8.16 we need to upper bound the quantity σ_n^t defined in (121). Recall the definition of $V_n^t(\sigma)$ from (120). By the above observations 1-3, we have

$$V_n^t(\sigma) \le 2q\left[\psi_n^\dagger(\sigma) + 4\sqrt{2}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\right] \tag{127}$$

We are only left to show that $\sigma_n^t := \inf\{\sigma : V_n^t(\sigma) \le 1\} \le \psi_n^\sharp(\frac{1}{4q}) + K\frac{t}{n}$, for a sufficiently large K, which will be implied if we can show that $V_n^t\left(\psi_n^\sharp(\frac{1}{2q}) + K\frac{t}{n}\right) \le 1$ (since then $\psi_n^\sharp(\frac{1}{2q}) + K\frac{t}{n} \in \{\sigma : V_n^t(\sigma) \le 1\}$ and the result follows from the minimality of σ_n^t). Note that, by the nonincreasing nature of each of the terms on the right hand side of (127),

$$\begin{aligned} V_n^t\left(\psi_n^\sharp(\frac{1}{4q}) + K\frac{t}{n}\right) &\le 2q\left[\psi_n^\dagger(\psi_n^\sharp(\frac{1}{4q})) + 4\sqrt{2}\sqrt{\frac{t}{n(Kt/n)}} + \frac{t}{n(Kt/n)}\right] \\ &\le 2q\left[\frac{1}{4q} + \frac{4\sqrt{2}}{\sqrt{K}} + \frac{1}{K}\right] < 1, \end{aligned}$$

where $K > 0$ is chosen so that $\frac{4\sqrt{2}}{\sqrt{K}} + \frac{1}{K} < \frac{1}{2}$ (note that $\psi_n^\dagger(\psi_n^\sharp(\frac{1}{4q})) \le \frac{1}{4q}$).

□

Example 8.19 (Finite dimensional classes)**.** *Suppose that* $\mathcal{L} \subset L_2(\Pi)$ *is a finite dimensional linear space with* $\dim(\mathcal{L}) = d < \infty$. *and let* $\mathcal{G} \subset \mathcal{L}$ *be a convex class of functions taking values in a bounded interval (for simplicity,* $[0, 1]$*). We would like to show that*

$$\mathbb{P}\left\{\|\hat{g}_n - g_*\|^2_{L_2(\Pi)} \ge \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + \left(\frac{d}{n} + K\frac{t}{n}\right)\right\} \le Ce^{-t} \tag{128}$$

with some constant $C, K > 0$.

It can be shown that[90] *that*

$$\psi_n(\delta) \le c\sqrt{\frac{d\delta}{n}}$$

with some constant $c > 0$. *Hence,*

$$\psi_n^\dagger(\sigma) = \sup_{\delta\ge\sigma} \frac{\psi_n(\delta)}{\delta} \le \sup_{\delta\ge\sigma} c\sqrt{\frac{d}{\delta n}} = c\sqrt{\frac{d}{\sigma n}}.$$

[89] Note that $\psi^\sharp$ can be thought of as the *generalized inverse* of $\psi^\dagger$. Thus, under the assumption that $\psi^\dagger$ is right-continuous, $\psi^\dagger(\sigma) \le \varepsilon$ if and only if $\sigma \ge \psi^\sharp(\varepsilon)$ (Exercise (HW3): Show this). Further note that with this notation $\sigma_n^t = V_n^{t,\sharp}(1)$.

[90] Exercise (HW3): Suppose that $\mathcal{L}$ is a finite dimensional subspace of $L_2(P)$ with $\dim(\mathcal{L}) = d$. Then $\mathbb{E}\left[\sup_{f\in\mathcal{L}:\|f\|_{L_2(P)}\le r} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \le r\sqrt{\frac{d}{n}}$.

As, $\psi_n^\dagger(\sigma) \le \varepsilon$ implies $\sigma \ge \psi_n^\sharp(\varepsilon)$, taking $\sigma := \frac{d}{n}$ and $q \ge \max\{1, 1/(4c)\}$, we see that

$$\psi_n^\dagger\left(\frac{d}{n}\right) \le c\sqrt{\frac{d}{\frac{d}{n}n}} \le \frac{1}{4q} \quad \Rightarrow \quad \psi_n^\sharp(\frac{1}{4q}) \le \frac{d}{n},$$

and Theorem 8.18 then implies (128); here $C \equiv C_q$ is taken as in Theorem 8.16 and K as in Theorem 8.18.

Exercise (HW3): Consider the setting of Example 8.19. Instead of using the refined analysis using (105) (and Talagrand's concentration inequality) as illustrated in this section, use the bounded differences inequality to get a crude upper bound on the excess risk of the ERM in this problem. Compare the obtained high probability bound to (128).

Exercise (HW3)[VC-subgraph classes]: Suppose that $\mathcal{G}$ is a convex VC-subgraph class of functions $g : \mathcal{Z} \to [0,1]$ of VC-dimension V. Then, show that, the function $\psi_n(\delta)$ can be upper bounded by:

$$\psi_n(\delta) \le c\left[\sqrt{\frac{V\delta}{n}\log\frac{1}{\delta}} \vee \frac{V}{n}\log\frac{1}{\delta}\right].$$

Show that $\psi_n^\sharp(\varepsilon) \le \frac{cV}{n\varepsilon^2}\log\frac{n\varepsilon^2}{V}$. Finally, use Theorem 8.18 to obtain a high probability bound analogous to (125).

Exercise (HW3)[Nonparametric classes]: In the case when the metric entropy of the class $\mathcal{G}$ (random, uniform, bracketing, etc.; e.g., if $\log N(\varepsilon, \mathcal{G}, L_2(\mathbb{P}_n)) \le \left(\frac{A}{\varepsilon}\right)^{2\rho}$) is bounded by $O(\varepsilon^{-2\rho})$ for some $\rho \in (0,1)$ (assuming that the envelope of $\mathcal{G}$ is 1), we typically have $\psi_n^\sharp(\varepsilon) \le O(n^{-1/(1+\rho)})$. Finally, use Theorem 8.18 to obtain a high probability bound analogous to (125).

8.4 Kernel density estimation

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. P on $\mathbb{R}^d$, $d \ge 1$. Suppose P has density p with respect to the Lebesgue measure on $\mathbb{R}^d$, and $\|p\|_\infty < \infty$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be any measurable function that integrates to one, i.e.,

$$\int_{\mathbb{R}^d} K(y)dy = 1$$

and $\|K\|_\infty < \infty$. Then the kernel density estimator (KDE) of p if given by

$$\widehat{p}_{n,h}(y) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{y - X_i}{h}\right) = h^{-d}\mathbb{P}_n\left[K\left(\frac{y-X}{h}\right)\right], \qquad \text{for } y \in \mathbb{R}^d.$$

Here h is called the smoothing bandwidth. Choosing a suitable bandwidth sequence $h_n \to 0$ and assuming that the density p is continuous, one can obtain a strongly consistent estimator $\widehat{p}_{n,h}(y) \equiv \widehat{p}_{n,h_n}(y)$ of $p(y)$, for any $y \in \mathbb{R}^d$.

It is natural to write the difference $\widehat{p}_n(y,h) - p(y)$ as the sum of a random term and a deterministic term:

$$\widehat{p}_{n,h}(y) - p(y) = \widehat{p}_{n,h}(y) - p_h(y) + p_h(y) - p(y)$$

where

$$p_h(y) := h^{-d} P\Big[K\Big(\frac{y-X}{h}\Big)\Big] = h^{-d}\int_{\mathbb{R}^d} K\Big(\frac{y-x}{h}\Big)p(x)dx = \int_{\mathbb{R}^d} K(u)p(y-hu)du$$

is a smoothed version of p. Convergence to zero of the second term can be argued based only on smoothness assumptions on p: if p is uniformly continuous, then it is easily seen that

$$\sup_{h\le b_n}\sup_{y\in\mathbb{R}^d} |p_h(y)-p(y)| \to 0$$

for any sequence $b_n \to 0$. On the other hand, the first term is just

$$h^{-d}(\mathbb{P}_n - P)\Big[K\Big(\frac{y-X}{h}\Big)\Big]. \tag{129}$$

For a fixed $y \in \mathbb{R}^d$, it is easy to study the properties of the above display using the CLT as we are dealing with a sum of independent random variables $h^{-d}K\Big(\frac{y-X_i}{h}\Big)$, $i = 1,\ldots,n$. However, it is natural to ask whether the KDE $\widehat{p}_{n,h_n}$ converges to p uniformly (a.s.) for a sequence of bandwidths $h_n \to 0$ and, if so, what is the rate of convergence in that case? We investigate this question using tools from empirical processes.

The KDE $\widehat{p}_{n,h}(\cdot)$ is indexed by the bandwidth h, and it is natural to consider $\widehat{p}_{n,h}$ as a process indexed by both $y \in \mathbb{R}^d$ and $h > 0$. This leads to studying the class of functions

$$\mathcal{F} := \left\{x \mapsto K\left(\frac{y-x}{h}\right) : y \in \mathbb{R}^d, h > 0\right\}.$$

It is fairly easy to give conditions on the kernel K so that the class $\mathcal{F}$ defined above satisfies

$$N(\epsilon\|K\|_\infty, \mathcal{F}, L_2(Q)) \le (A/\epsilon)^V \tag{130}$$

for some constants $V \ge 2$ and $A \ge e^2$; see e.g., Lemma 7.22[91]. While it follows immediately from the GC theorem that

$$\sup_{h>0, y\in\mathbb{R}^d}\left|(\mathbb{P}_n - P)\Big[K\Big(\frac{y-X}{h}\Big)\Big]\right| \overset{a.s.}{\to} 0,$$

this does not suffice in view of the factor of h^{-d} in (129). In fact, we need a rate of convergence for $\sup_{h>0,y\in\mathbb{R}^d}(\mathbb{P}_n - P)\Big[K\Big(\frac{y-X}{h}\Big)\Big] \overset{a.s.}{\to} 0$. The following theorem gives such a result[92].

[91] For instance, it is satisfied for general $d \ge 1$ whenever $K(x) = \phi(q(x))$, with $q(x)$ being a polynomial in d variables and ϕ being a real-valued right continuous function of bounded variation.

[92] To study variable bandwidth kernel estimators [Einmahl and Mason, 2005] derived the following result, which can be proved with some extra effort using ideas from the proof of Theorem 8.21.

Theorem 8.20. *For any $c > 0$, with probability 1,*

$$\limsup_{n\to\infty}\sup_{c\log n/n \le h \le 1} \frac{\sqrt{nh}\|\widehat{p}_{n,h}(y) - p_h(y)\|_\infty}{\sqrt{\log(1/h)\vee\log\log n}} =: K(c) < \infty.$$

Theorem (8.20) implies for any sequences $0 < a_n < b_n \le 1$, satisfying $b_n \to 0$ and $na_n/\log n \to \infty$, with probability 1 ,

$$\sup_{a_n\le h\le b_n}\|\widehat{p}_{n,h} - p_h\|_\infty = O\left(\sqrt{\frac{\log(1/a_n)\vee\log\log n}{na_n}}\right),$$

which in turn implies that $\lim_{n\to\infty}\sup_{a_n\le h\le b_n}\|\widehat{p}_{n,h} - p_h\|_\infty \overset{a.s.}{\to} 0$.

Theorem 8.21. *Suppose that $h_n \downarrow 0$, $nh_n^d/|\log h_n| \to \infty$, $\log\log n/|\log h_n| \to \infty$ and $h_n^d \le \check{c}h_{2n}^d$ for some $\check{c} > 0$. Then*

$$\limsup_{n\to\infty} \frac{\sqrt{nh_n^d}\|\widehat{p}_{n,h_n}(\cdot) - p_{h_n}(\cdot)\|_\infty}{\sqrt{\log h_n^{-1}}} = C \quad a.s.$$

where $C < \infty$ is a constant that depends only on the VC characteristics of $\mathcal{F}$.

Proof. We will use the following result:

Lemma 8.22 ([de la Peña and Giné, 1999, Theorem 1.1.5])**.** *If $X_i, i \in \mathbb{N}$, are i.i.d $\mathcal{X}$-valued random variables and $\mathcal{F}$ a class of measurable functions, then*

$$\mathbb{P}\left(\max_{1\le j\le n}\Big\|\sum_{i=1}^{j}(f(X_i) - Pf)\Big\|_{\mathcal{F}} > t\right) \le 9\,\mathbb{P}\Big(\Big\|\sum_{i=1}^{n}(f(X_i) - Pf)\Big\|_{\mathcal{F}} > \frac{t}{30}\Big).$$

For $k \ge 0$, let $n_k := 2^k$. Let $\lambda > 0$; to be chosen later. The monotonicity of $\{h_n\}$ (hence of $h_n \log h_n^{-1}$ once $h_n < e^{-1}$) and Lemma 8.22 imply (for $k \ge 1$)

$$\begin{aligned}
& \mathbb{P}\left(\max_{n_{k-1}<n\le n_k}\sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h_n}(y) - p_{h_n}(y)\|_\infty > \lambda\right)\\
= \;& \mathbb{P}\left(\max_{n_{k-1}<n\le n_k}\sqrt{\frac{1}{nh_n^d\log h_n^{-1}}}\sup_{y\in\mathbb{R}^d}\left|\sum_{i=1}^{n}\Big[K\Big(\frac{y-X_i}{h_n}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h_n}\Big)\Big]\right| > \lambda\right)\\
\le \;& \mathbb{P}\left(\frac{1}{\sqrt{n_{k-1}h_{n_k}^d\log h_{n_k}^{-1}}} \times \max_{1\le n\le n_k}\sup_{y\in\mathbb{R}^d, h_{n_k}\le h\le h_{n_{k-1}}}\left|\sum_{i=1}^{n}\Big[K\Big(\frac{y-X_i}{h}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h}\Big)\Big]\right| > \lambda\right)\\
\le \;& 9\mathbb{P}\left(\frac{1}{\sqrt{n_{k-1}h_{n_k}^d\log h_{n_k}^{-1}}} \times \sup_{y\in\mathbb{R}^d, h_{n_k}\le h\le h_{n_{k-1}}}\left|\sum_{i=1}^{n_k}\Big[K\Big(\frac{y-X_i}{h}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h}\Big)\Big]\right| > \frac{\lambda}{30}\right).
\end{aligned} \tag{131}$$

We will study the subclasses

$$\mathcal{F}_k := \left\{K\Big(\frac{y-\cdot}{h}\Big) : h_{n_k} \le h \le h_{n_{k-1}}, y \in \mathbb{R}^d\right\}.$$

As

$$\mathbb{E}\Big[K^2\Big(\frac{y-X}{h}\Big)\Big] = \int_{\mathbb{R}^d} K^2\Big(\frac{y-x}{h}\Big)p(x)dx = h^d\int_{\mathbb{R}^d} K^2(u)p(y-uh)du \le h^d\|p\|_\infty\|K\|_2^2,$$

for the class $\mathcal{F}_k$, we can take

$$U_k := 2\|K\|_\infty, \qquad \text{and} \qquad \sigma_k^2 := h_{n_{k-1}}^d\|p\|_\infty\|K\|_2^2.$$

Since $h_{n_k} \downarrow 0$, and $nh_n^d/\log h_n^{-1} \to \infty$, there exists $k_0 < \infty$ such that for all $k \ge k_0$,

$$\sigma_k < U_k/2 \qquad \text{and} \qquad \sqrt{n_k}\sigma_k \ge \sqrt{V}U_k\sqrt{\log\frac{AU_k}{\sigma_k}}. \qquad \text{(check!)} \tag{132}$$

Letting $Z_k := \mathbb{E}\Big\| \sum_{i=1}^{n_k}(f(X_i) - Pf)\Big\|_{\mathcal{F}_k}$, we can bound $\mathbb{E}[Z_k]$ by using Theorem 7.13 (see (84)), for $k \geq k_0$, to obtain

$$\mathbb{E}[Z_k] = \mathbb{E}\Big\| \sum_{i=1}^{n_k}(f(X_i) - Pf)\Big\|_{\mathcal{F}_k} \leq L\sigma_k\sqrt{n_k \log(AU_k/\sigma_k)}$$

for a suitable constant $L > 0$. Thus, using (132),

$$\nu_k := n_k\sigma_k^2 + 2U_k\mathbb{E}[Z_k] \leq \tilde{c}n_k\sigma_k^2$$

for a constant $\tilde{c} > 1$ and $k \geq k_0$. Choosing $x = c\log(AU_k/\sigma_k)$ in (99), for some $c > 0$, we see that

$$\begin{aligned}\mathbb{E}[Z_k] + \sqrt{2\nu_k x} + U_k x/3 &\leq \sigma_k\sqrt{n_k \log(AU_k/\sigma_k)}(L + \sqrt{2c\tilde{c}}) + cU_k\log(AU_k/\sigma_k)/3 \\ &\leq C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)},\end{aligned}$$

for some constant $C > 0$, where we have again used (132). Therefore, by Theorem 8.7,

$$\mathbb{P}\Big(Z_k \geq C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}\Big) \leq \mathbb{P}(Z_k \geq \mathbb{E}[Z_k] + \sqrt{2\nu_k x} + U_k x/3) \leq e^{-c\log(AU_k/\sigma_k)}.$$

Notice that

$$\frac{30C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}}{\sqrt{n_{k-1}h_{n_k}^d \log h_{n_k}^{-1}}} > \lambda \qquad \text{(check!)}$$

for some $\lambda > 0$, not depending on k. Therefore, choosing this λ the probability on the right hand-side of (131) can be expressed as

$$\mathbb{P}\left(\frac{Z_k}{\sqrt{n_{k-1}h_{n_k}^d \log h_{n_k}^{-1}}} > \frac{\lambda}{30}\right) \leq \mathbb{P}\Big(Z_k \geq C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}\Big) \leq e^{-c\log(AU_k/\sigma_k)}.$$

Since

$$\sum_{k=k_0}^{\infty} e^{-c\log(AU_k/\sigma_k)} = c_1\sum_{k=k_0}^{\infty} h_{n_{k-1}}^{cd/2} \leq \tilde{c}_1\sum_{k=k_0}^{\infty}(\check{c})^{-cd/2} < \infty,$$

for constants $c_1, \tilde{c}_1 > 0$, we get, summarizing,

$$\sum_{k=1}^{\infty}\mathbb{P}\left(\max_{n_{k-1}<n\leq n_k}\sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h}(y) - p_h(y)\|_\infty > \lambda\right) < \infty.$$

Let $Y_n = \sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h} - p_h\|_\infty$. Letting $Y := \limsup_{n\to\infty} Y_n$, and using the Borel-Cantelli lemma we can see that $\mathbb{P}(Y > \lambda) = 0$. This yields the desired result using the zero-one law[93]. □

[93]For a fixed $\lambda \geq 0$, define the event $A := \{\limsup_{n\to\infty} Y_n > \lambda\}$. As this is a tail event, by the zero-one law it has probability 0 or 1. We thus have that for each λ, $\mathbb{P}(Y > \lambda) \in \{0, 1\}$. Defining $c := \sup\{\lambda : \mathbb{P}(Y > \lambda) = 1\}$, we get that $Y = c$ a.s. Note that $c < \infty$ as there exists $\lambda > 0$ such that $\mathbb{P}(Y > \lambda) = 0$, by the proof of Theorem 8.21.

9 Review of weak convergence in complete separable metric spaces

In this chapter (and the next few) we will study weak convergence of stochastic processes. Suppose that $U_1, \ldots, U_n$ are i.i.d. Uniform$(0,1)$ random variables (i.e., U_1 has distribution function $G(x) = x$, for $x \in [0,1]$) and let G_n denote the empirical distribution function of $U_1, \ldots, U_n$. In this chapter, we will try to make sense of the (informal) statement:

$$\sqrt{n}(G_n - G) \stackrel{d}{\to} \mathbb{G}, \qquad \text{as } n \to \infty,$$

where $\mathbb{G}$ is a standard Brownian bridge process in $[0,1]$.

This will give us the right background and understanding to appreciate the Donsker theorem for the empirical process (indexed by an arbitrary class of functions). As we have seen in Chapter 1, weak convergence coupled with the continuous mapping theorem can help us study (in a unified fashion) the limiting distribution of many random variables that can be expressed as functionals of the empirical process.

9.1 Weak convergence of random vectors in $\mathbb{R}^d$

Before we study weak convergence of stochastic processes let us briefly recall the notion of weak convergence of random vectors.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let X be a random vector (or a measurable map) in $\mathbb{R}^k$ $(k \geq 1)$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$[94].

Definition 9.1. *Let $\{X_n\}_{n\geq 1}$ be a sequence of random vectors in $\mathbb{R}^k$. Let X_n have distribution function*[95] *F_n and let P_n denote the distribution*[96] *of X_n, i.e., $X_n \sim P_n$. We say that $\{X_n\}$ converges in distribution (or weakly or in law) to a random vector $X \sim P$ (or to the distribution P) with distribution function F if*

$$F_n(x) \to F(x), \qquad \text{as } n \to \infty, \tag{133}$$

for every $x \in \mathbb{R}^k$ at which the limiting distribution function $F(\cdot)$ is continuous. This is usually denoted by $X_n \stackrel{d}{\to} X$ or $P_n \stackrel{d}{\to} P$.

As the name suggests, weak convergence only depends on the induced laws of the vectors and not on the probability spaces on which they are defined.

The whole point of weak convergence is to approximate the probabilities of events related to X_n (i.e., P_n), for n large, by that of the limiting random vector X (i.e., P).

[94] More formally, $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^k, \mathcal{B}_k)$, where $\mathcal{B}_k$ is the Borel σ-field in $\mathbb{R}^d$, is a map such that for any $B \subset \mathcal{B}_k$, $X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$.

[95] Thus, $F_n(x) = \mathbb{P}(X_n \leq x)$, for all $x \in \mathbb{R}^d$.

[96] Thus, $P_n : \mathcal{B}_k \to [0,1]$ is a map such that $P_n(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$ for all $B \in \mathcal{B}_k$.

The multivariate central limit theorem (CLT) is a classic application of convergence in distribution and it highlights the usefulness of such a concept.

The following theorem, referred to as the Portmanteau result, gives a number of equivalent descriptions of weak convergence (most of these characterizations are only useful in proofs). Indeed, the characterization (v) (of the following result) makes the intuitive notion of convergence in distribution rigorous.

Theorem 9.2 (Portmanteau theorem). *Let X, $X_1, \ldots, X_n$ be random vectors in $\mathbb{R}^k$. Let $X_n \sim P_n$, for $n \geq 1$, and $X \sim P$. Then, the following are equivalent:*

(i) $X_n \stackrel{d}{\to} X$ *or* $P_n \stackrel{d}{\to} P$, *i.e.,* (133) *holds;*

(ii) $\int f dP_n =: P_n f \to Pf := \int f dP$, *as $n \to \infty$, for every $f : \mathbb{R}^k \to \mathbb{R}$ which is continuous and bounded.*

(iii) $\liminf_{n\to\infty} P_n(G) \geq P(G)$ *for all open $G \subset \mathbb{R}^k$;*

(iv) $\limsup_{n\to\infty} P_n(F) \leq P(F)$ *for all closed $F \subset \mathbb{R}^k$;*

(v) $P_n(A) \to P(A)$ *for all P-continuity sets A (i.e., $P(\partial A) = 0$, where ∂A denotes the boundary of A).*

Proof. See Lemma 2.2 of [van der Vaart, 1998]. □

Quite often we are interested in the distribution of a one-dimensional (or m-dimensional, where $m \leq k$ usually) function of X_n. The following result, the continuous mapping theorem is extremely useful and, in some sense, justifies the study of weak convergence.

Theorem 9.3 (Continuous mapping). *Suppose that $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. If $X_n \stackrel{d}{\to} X$ then $g(X_n) \stackrel{d}{\to} g(X)$.*

9.2 Weak convergence in metric spaces and the continuous mapping theorem

As before, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and and let $(T, \mathcal{T})$ is a measurable (metric) space. $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (T, \mathcal{T})$ is a *random element* if $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{T}$. Quite often, when we talk about a random element X, we take $\mathcal{T} = \mathcal{B}(T)$, the Borel σ-field (on the set T), the smallest σ-field containing all open sets.

Question: How do we define the notion of weak convergence in this setup?

Although it is not straight forward to define weak convergence as in Definition 9.1 (for random vectors through their distribution functions), the equivalent definition (ii) in Theorem 9.2 can be extended easily.

Let $\mathcal{C}_b(T; \mathcal{T})$ denote the set of all *bounded, continuous*, $\mathcal{T}/\mathcal{B}(\mathbb{R})$-measurable, real-valued functions on T.

Definition 9.4. *A sequence* $\{X_n\}_{n\geq 1}$ *of random elements of* T *(defined on* $(\Omega, \mathcal{A}, \mathbb{P})$*) converges in distribution to a random element* X*, written* $X_n \stackrel{d}{\to} X$*, if and only if*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X) \qquad \text{for each} \qquad f \in \mathcal{C}_b(T;\mathcal{T}). \tag{134}$$

Similarly, a sequence $\{P_n\}_{n\geq 1}$ *of probability measures on* $\mathcal{T}$ *converges weakly to* P*, written* $P_n \stackrel{d}{\to} P$*, if and only if*

$$P_n f \to P f \qquad \text{for each} \qquad f \in \mathcal{C}_b(T;\mathcal{T}), \tag{135}$$

where $P_n f := \int f dP_n$ *and* $Pf := \int f dP$*.*

Note that definition (135) of weak convergence is slightly more general than (134) as it allows the X_n's to be defined on different probability spaces (and we let $X_n \sim P_n$, for $n \geq 1$, and $X \sim P$). As in the previous subsection, we have the following equivalent characterizations of weak convergence.

9.2.1 When $\mathcal{T} = \mathcal{B}(T)$, the Borel σ-field of T

Suppose that T is a metric space with metric $d(\cdot,\cdot)$ and let $\mathcal{T}$ be the Borel σ-field of T generated by the metric $d(\cdot,\cdot)$. Then, we have the following equivalent characterizations of weak convergence.

Theorem 9.5. *(Portmanteau theorem) Let* $X, X_1, \ldots, X_n$ *be random elements taking values in a metric space* $(T, \mathcal{B}(T))$*. Let* $X_n \sim P_n$*, for* $n \geq 1$*, and* $X \sim P$*. Then, the following are equivalent:*

(i) $X_n \stackrel{d}{\to} X$ *or* $P_n \stackrel{d}{\to} P$*;*

(ii) $\liminf_{n\to\infty} P_n(G) \geq P(G)$ *for all open* $G \subset T$*;*

(iii) $\limsup_{n\to\infty} P_n(F) \leq P(F)$ *for all closed* $F \subset T$*;*

(iv) $P_n(A) \to P(A)$ *for all* P*-continuity sets* A *(i.e.,* $P(\partial A) = 0$*, where* ∂A *denotes the boundary of* A*).*

Proof. See Theorem 2.1 of [Billingsley, 1999] (or 3.25 of [Kallenberg, 2002]). □

Such an abstract definition of weak convergence can only be useful if we have a continuous mapping theorem (as before). The following result shows that essentially the "vanilla' version of the continuous mapping theorem is true, and is a trivial consequence of the definition of weak convergence.

Theorem 9.6 (Continuous mapping theorem)**.** *Let $(T, \mathcal{B}(T))$ be a measurable (metric) space. Suppose that $\{X_n\}_{n\geq 1}$ is a sequence of random elements of T converging in distribution to a random element X, i.e., $X_n \xrightarrow{d} X$. Let $H : (T, \mathcal{B}(T)) \to (S, \mathcal{B}(S))$ be a continuous function, where $(S, \mathcal{B}(S))$ is another measurable (metric) space. Then $H(X_n) \xrightarrow{d} H(X)$ as random elements in S.*

Proof. Let $f \in \mathcal{C}_b(S, \mathcal{B}(S))$. Then $f \circ H \in \mathcal{C}_b(T, \mathcal{B}(T))$ and thus from the definition of weak convergence of X_n,

$$\mathbb{E}[(f \circ H)(X_n)] \to \mathbb{E}[(f \circ H)(X)].$$

As the above convergence holds for all $f \in \mathcal{C}_b(S, \mathcal{B}(S))$, again using the definition of weak convergence, we can say that $H(X_n) \xrightarrow{d} H(X)$. $\square$

Exercise(HW4): Show that for any sequence of points $y_n \in S$, $n \geq 1$, (where S is a metric space with Borel σ-field $\mathcal{B}(S)$) $y_n \to y_0$ if and only if $\delta_{y_n} \xrightarrow{d} \delta_{y_0}$, where $\delta.$ is the Dirac delta measure.

As we will see later, requiring that H be continuous of the entire set T is asking for too much. The following result, which can be thought of as an extension of the above result (and is similar in flavor to Theorem 9.3), requires that H be continuous on the set where the limiting random element X lives (or on the support of the limiting probability measure).

Theorem 9.7. *Let $H : (T, \mathcal{B}(T)) \to (S, \mathcal{B}(S))$ be a measurable mapping. Write C for the set of points in T at which H is continuous. If $X_n \xrightarrow{d} X$ for which $\mathbb{P}(X \in C) = 1$, then $H(X_n) \xrightarrow{d} H(X)$.*

Proof. Let $X_n \sim P_n$ and $X \sim P$. Fix $\psi \in \mathcal{C}_b(S; \mathcal{B}(S))$. Then the measurable, real-valued, bounded function $h := \psi \circ H$ is continuous at all points in C. We will have to show that $P_n h \to Ph$ as $n \to \infty$.

Consider any increasing sequence $\{f_i\}$ of bounded, continuous functions for which $f_i \leq h$ everywhere and $f_i \uparrow h$ at each point of C. Accept for the moment that such a sequence exists. Then, weak convergence of P_n to P implies that

$$P_n f_i \to P f_i \qquad \text{for each fixed } f_i.$$

Thus,

$$\liminf_{n\to\infty} P_n h \geq \liminf_{n\to\infty} P_n f_i = P f_i \qquad \text{for each fixed } f_i.$$

Invoking monotone convergence as $i \to \infty$ on the right-hand side yields

$$\liminf_{n\to\infty} P_n h \geq \lim_{i\to\infty} P f_i = Ph.$$

By substituting $-h$ for h and going through the above argument again gives the desired result.

It only remains to show that we can construct such a sequence $\{f_i\}$. These functions must be chosen from the family

$$\mathcal{F} = \{f \in \mathcal{C}_b(T; \mathcal{B}(T)) : f \leq h\}.$$

If we can find a countable subfamily of $\mathcal{F}$, say $\{g_1, g_2, \ldots\}$, whose pointwise supremum equals h at each point of C, then setting $f_i := \max\{g_1, \ldots, g_i\}$ will do the trick.

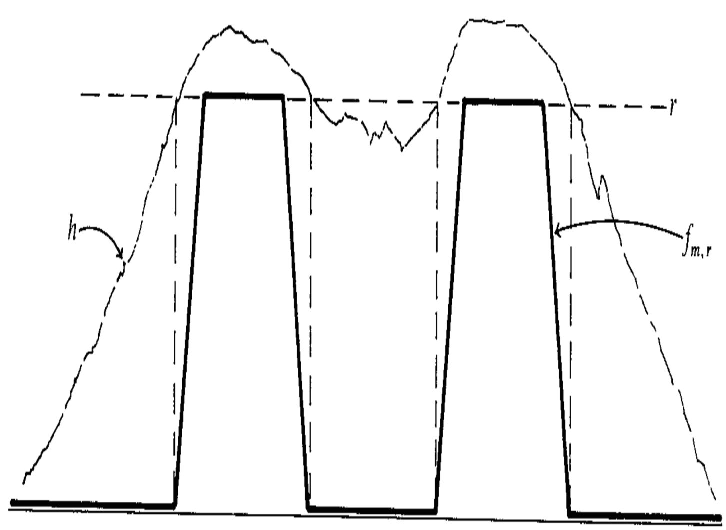
Without loss of generality suppose that $h > 0$ (a constant could be added to h to achieve this). Let $d(\cdot, \cdot)$ be the metric in T. For each subset $A \in \mathcal{B}(T)$, define the distance function $d(\cdot, A)$ by

$$d(x, A) = \inf\{d(x, y) : y \in A\}.$$

Note that $d(\cdot, A)$ is a continuous function (in fact, $d(\cdot, A)$ is uniformly continuous. Exercise (HW4): Show this.), for each fixed A. For a positive integer m and a positive rational r define

$$f_{m,r}(x) := r \wedge [m\, d(x, \{h \leq r\})].$$

Each $f_{m,r}$ is bounded and continuous; it is at most r if $h(x) > r$; it takes the value zero if $h(x) \leq r$. Thus, $f_{m,r} \in \mathcal{F}$.



Given a point $x \in C$ and an $\epsilon > 0$, choose a positive rational number r with $h(x) - \epsilon < r < h(x)$. Continuity of h at x keeps its value greater than r in some neighborhood of x. Consequently, $d(x, \{h \leq r\}) > 0$ and $f_{m,r}(x) = r > h(x) - \epsilon$ for all m large enough. Take $\{f_{m,r} : m \in \mathbb{N}_+, r \in \mathbb{Q}_+\}$ as the countable set $\{g_1, g_2, \ldots\}$. $\square$

9.2.2 The general continuous mapping theorem

Till now we have restricted our attention to metric spaces $(T, \mathcal{T})$ endowed with their Borel σ-field, i.e., $\mathcal{T} = \mathcal{B}(\mathcal{T})$. In this subsection we extend the continuous mapping theorem when

$\mathcal{T}$ need not equal $\mathcal{B}(T)$.

Formally, we ask the following question. Suppose that $X_n \xrightarrow{d} X$ as $\mathcal{T}$-measurable random elements of a metric space T, and let H be a $\mathcal{T}/\mathcal{S}$-measurable map from T into another metric space S. If H is continuous at each point of an $\mathcal{T}$-measurable set C with $\mathbb{P}(X \in C) = 1$, does it follow that $H(X_n) \xrightarrow{d} H(X)$, i.e., does $\mathbb{E}f(H(X_n))$ converge to $\mathbb{E}f(H(X))$ for every $f \in \mathcal{C}_b(S;\mathcal{S})$?

We will now assume that $\mathcal{T}$ is a sub-σ-field of $\mathcal{B}(T)$. Define

$$\mathcal{F} = \{f \in \mathcal{C}_b(T;\mathcal{T}) : f \leq h\}.$$

Last time we constructed the countable subfamily that took the form

$$f_{m,r}(x) = r \wedge [md(x, \{h \leq r\}].$$

Continuity of $f_{m,r}$ suffices for Borel measurability, but it needn't imply $\mathcal{T}$-measurability. We must find a substitute for these functions.

Definition 9.8. *Call a point $x \in T$ completely regular (with respect to the metric d and the σ-field $\mathcal{T}$) if to each neighborhood V of x there exists a uniformly continuous, $\mathcal{T}$-measurable function g with $g(x) = 1$ and $g \leq \mathbf{1}_V$.*

Theorem 9.9 (Continuous mapping theorem)**.** *Let $(T,\mathcal{T})$ is a measurable (metric) space. Suppose that $\{X_n\}_{n\geq 1}$ is a sequence of random elements of T converges in distribution to a random element X, i.e., $X_n \xrightarrow{d} X$. Let $H : (T,\mathcal{T}) \to (S,\mathcal{S})$ be a continuous function at each point of some separable, $\mathcal{T}$-measurable set C of completely regular points such that $\mathbb{P}(X \in C) = 1$, then $H(X_n) \xrightarrow{d} H(X)$ as random elements in $(S,\mathcal{S})$.*

Idea of proof: (Step 1) Use the fact that each $x \in C$ is completely regular to approximate $h(x)$ by a supremum of a sub-class of $\mathcal{C}_b(T;\mathcal{T})$ that is uniformly continuous. (Step 2) Then use the separability of C to find a sequence of uniformly continuous functions in $\mathcal{C}_b(T;\mathcal{T})$ that increase to h for all $x \in C$. (Step 3) Use the same technique as in the initial part of the proof of Theorem 9.7 to complete the proof.

Proof. Let $X_n \sim P_n$ and $X \sim P$. Fix $\psi \in C_b(S;\mathcal{B}(S))$. Then the $\mathcal{T}$-measurable, real-valued, bounded function $h := \psi \circ H$ is continuous at all points in C. We will have to show that $P_n h \to Ph$.

Without loss of generality suppose that $h > 0$ (a constant could be added to h to achieve this). Define

$$\mathcal{F} = \{f \in \mathcal{C}_b(T;\mathcal{T}) : f \leq h; f \text{ is uniformly continuous}\}.$$

At those completely regular points x of T where h is continuous, the supremum of $\mathcal{F}$ equals h, i.e.,

$$h(x) = \sup_{f\in\mathcal{F}} f(x). \tag{136}$$

To see this suppose that h is continuous at a point x that is completely regular. Choose r with $0 < r < h(x)$ (r should ideally be close to $h(x)$). By continuity, there exists a $\delta > 0$ such that $h(y) > r$ on the closed ball $B(x, \delta)$. As x is completely regular, there exists a uniformly continuous, $\mathcal{T}$-measurable g such that $g(x) = 1$ and $g \leq \mathbf{1}_{B(x,\delta)}$. Now, look at the function $f = rg$. Observe that $f \in \mathcal{F}$ and $f(x) = r$. Thus (136) holds for all $x \in C$.

Separability of C will enable us to extract a suitable countable subfamily from $\mathcal{F}$. Let C_0 be a countable dense subset of C. Let $\{g_1, g_2, \ldots\}$ be the set of all those functions of the form $r\mathbf{1}_B$, with r rational, B a closed ball of rational radius centered at a point of C_0, and $r\mathbf{1}_B \leq f$ for at least one f in $\mathcal{F}$. For each g_i choose one f satisfying the inequality $g_i \leq f$. Denote it by f_i. This picks out the required countable subfamily:

$$\sup_i f_i = \sup_{f \in \mathcal{F}} f \qquad \text{on } C. \tag{137}$$

To see this, consider any point $z \in C$ and any $f \in \mathcal{F}$. For each rational number r such that $f(z) > r > 0$ choose a rational ϵ for which $f > r$ at all points within a distance 2ϵ of z. Let B be the closed ball of radius ϵ centered at a point $x \in C_0$ for which $d(x, z) < \epsilon$. The function $r\mathbf{1}_B$ lies completely below f; it must be one of the g_i. The corresponding f_i takes a value greater than r at z. Thus, assertion (137) follows.

The proof can now be completed using analogous steps as in the proof of Theorem 9.7. Assume without loss of generality that $f_i \uparrow h$ at each point of C. Then,

$$\begin{aligned} \liminf_{n\to\infty} P_n h &\geq \liminf_{n\to\infty} P_n f_i && \text{for each } i \\ &= P f_i && \text{because } P_n \xrightarrow{d} P \\ &\to P h && \text{as } i \to \infty, \text{ by monotone convergence.} \end{aligned}$$

Replace h by $-h$ (plus a big constant) to get the companion inequality for the $\limsup$. $\square$

Corollary 9.10. *If $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for each bounded, uniformly continuous, $\mathcal{T}$-measurable f, and if X concentrates on a separable set of completely regular points, then $X_n \xrightarrow{d} X$.*

The corollary follows directly from the decision to insist upon uniform continuous separating functions in the definition of a completely regular point.

9.3 Weak convergence in the space $C[0, 1]$

Till now we have mostly confined ourselves to finite dimensional spaces (although the discussion in the previous subsection is valid for any metric space). In this subsection we briefly discuss an important, both historically and otherwise, infinite dimensional metric space where our random elements take values. Let $C[0, 1]$ denote the *space of continuous functions* from $[0, 1]$ to $\mathbb{R}$. We endow $C[0, 1]$ with the *uniform metric*:

$$\rho(x, y) := \|x - y\|_\infty := \sup_{t \in [0,1]} |x(t) - y(t)|, \qquad \text{for } x, y \in C[0, 1].$$

Remark 9.1. *Note that $C[0,1]$, equipped with the uniform metric, is separable*[97] *(Exercise (HW4): Show this.) and complete*[98]*. $C[0,1]$ is complete because a uniform limit of continuous functions is continuous [Browder, 1996, Theorem 3.24]. $C[0,1]$ is separable because the set of polynomials on $[0,1]$ is dense in $C[0,1]$ by the Weierstrass approximation theorem [Browder, 1996, Theorem 7.1], and the set of polynomials with rational coefficients is countable and dense in the set of all polynomials, hence also dense in $C[0,1]$.*

Question: Let X and X_n are random processes (elements) on $[0,1]$. When can we say that $X_n \stackrel{d}{\to} X$?

Suppose that $X_n \stackrel{d}{\to} X$ in $C[0,1]$. For each $t \in [0,1]$ define the *projection* map $\pi_t : C[0,1] \to \mathbb{R}$ as

$$\pi_t(x) = x(t), \qquad \text{for } x \in C[0,1].$$

Observe that π_t is a uniformly continuous map as $|\pi_t(x) - \pi_t(y)| = |x(t) - y(t)| \le \|x - y\|_\infty$ for all $x, y \in C[0,1]$. Thus, by the continuous mapping theorem (see e.g., Theorem 9.7), we know that $X_n(t) = \pi_t(X_n) \stackrel{d}{\to} \pi_t(X) = X(t)$ for all $t \in [0,1]$.

We shall write $X_n \stackrel{fd}{\to} X$ for *convergence of the finite-dimensional distributions*, in the sense that

$$(X_n(t_1), \dots, X_n(t_k)) \stackrel{d}{\to} (X(t_1), \dots, X(t_k)), \qquad t_1, \dots, t_k \in [0,1],\ k \in \mathbb{N}. \tag{138}$$

Indeed, by defining the projection map $\pi_{(t_1,\dots,t_k)} : C[0,1] \to \mathbb{R}^k$ as $\pi_{(t_1,\dots,t_k)}(x) = (x(t_1), \dots, (t_k))$, for $t_1, \dots, t_k \in [0,1]$, we can show that if $X_n \stackrel{d}{\to} X$ in $C[0,1]$ then $X_n \stackrel{fd}{\to} X$.

Remark 9.2. *Although it can be shown that the distribution of a random process is determined by the family of finite-dimensional distributions*[99]*, condition (138) is insufficient in general for the convergence $X_n \stackrel{d}{\to} X$. Hint: Consider the sequence of points $x_n \in C[0,1]$, for $n \ge 1$, where*

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

Let $x_0 \equiv 0 \in C[0,1]$. We can show that $\delta_{x_n} \stackrel{fd}{\to} \delta_{x_0}$. But δ_{x_n} does NOT converge weakly to δ_{x_0} as $x_n \not\to x_0$.

[97] A topological space is *separable* if it contains a countable dense subset; that is, there exists a sequence $\{x_n\}_{n=1}^\infty$ of elements of the space such that every nonempty open subset of the space contains at least one element of the sequence.

[98] A metric space T is called *complete* (or a Cauchy space) if every Cauchy sequence of points in T has a limit that is also in T or, alternatively, if every Cauchy sequence in T converges in T. Intuitively, a space is complete if there are no "points missing" from it (inside or at the boundary).

[99] We have the following result (a consequence of [Kallenberg, 2002, Proposition 3.2]; also see [Billingsley, 1999, Chapter 1]).

Lemma 9.11. *Suppose that X and Y are random elements in $C[0,1]$. Then $X \stackrel{d}{=} Y$ if and only if*

$$(X(t_1), \dots, X(t_k)) \stackrel{d}{=} (Y(t_1), \dots, Y(t_k)), \qquad t_1, \dots, t_k \in [0,1],\ k \in \mathbb{N}.$$

The above result holds more generally than $C[0,1]$.

Alternatively, if $\delta_{x_n} \xrightarrow{d} \delta_{x_0}$*, then* $\|x_n\|_\infty := \sup_{t\in[0,1]} x_n(t) = 1$ *should converge weakly to* $\|x_0\|_\infty = 0$ *(which is obviously not true!). Note that here we have used the fact that* $f : C[0,1] \to \mathbb{R}$*, defined as* $f(x) = \sup_{t\in[0,1]} x(t)$ *is a continuous function (Exercise(HW4): Show this.).*

The above discussion motivates the following concepts, which we will show are crucially tied to the notion of weak convergence of stochastic processes.

9.3.1 Tightness and relative compactness

Suppose that X_n's are random elements taking values in the metric space $(T, \mathcal{B}(T))$.

Definition 9.12. *We say that the sequence of random elements* $\{X_n\}_{n\geq 1}$ *is tight if and only if for every* $\epsilon > 0$ *there exists a compact set*[100] $V_\epsilon \subset T$ *such that*

$$\mathbb{P}(X_n \in V_\epsilon) > 1 - \epsilon, \qquad \textit{for all } n \geq 1.$$

Exercise (HW4): Let S be a *separable* and *complete* metric space. Then every probability measure on $(S, \mathcal{B}(S))$ is tight. (Hint: See [Billingsley, 1999, Theorem 1.3])

Recall that a set A is *relatively compact* if its closure is compact, which is equivalent to the condition that each sequence in A contains a convergent subsequence (the limit of which may not lie in A). This motivates the following definition which can be thought of as a 'probabilistic' version of relative compactness.

Definition 9.13. *A sequence of random elements* $\{X_n\}_{n\geq 1}$ *is said to be relatively compact in distribution if every subsequence has a further sequence that converges in distribution. Similarly, we can define the notion of relative compactness (in distribution) of a sequence* $\{P_n\}_{n\geq 1}$ *of probability measures.*

Theorem 9.14 (Prohorov's theorem)**.** *If* $\{X_n\}_{n\geq 1}$ *is tight, then it is relatively compact in distribution. In fact, the two notions are equivalent if* T *is separable and complete*[101]*.*

Proof. See Theorem 2.1 of [Billingsley, 1999] (or 3.25 of [Kallenberg, 2002]). □

Prohorov's theorem is probably the key result in this theory of classical weak convergence and gives the basic connection between tightness and relative distributional compactness.

[100] Recall that in a metric space a set is *compact* if and only if it is complete and totally bounded.

[101] A metric space M is called *complete* if every Cauchy sequence of points in M has a limit that is also in M; or, alternatively, if every Cauchy sequence in M converges in M.

9.3.2 Tightness and weak convergence in $C[0,1]$

Lemma 9.15. *Let $X, X_1, \ldots, X_n, \ldots$ be random elements in $C[0,1]$. Then $X_n \xrightarrow{d} X$ if and only if $X_n \xrightarrow{fd} X$ and $\{X_n\}$ is relatively compact in distribution.*

Proof. If $X_n \xrightarrow{d} X$, then $X_n \xrightarrow{fd} X$ (by the continuous mapping theorem; take $f : C[0,1] \to \mathbb{R}^k$ defined as $f(x) = (x(t_1), \ldots, x(t_k)))$, and $\{X_n\}$ is trivially relatively compact in distribution.

Now assume that $\{X_n\}$ satisfies the two conditions. If $X_n \stackrel{d}{\nrightarrow} X$, we may choose a bounded continuous function $f : C[0,1] \to \mathbb{R}$ and an $\epsilon > 0$ such that $|\mathbb{E}f(X_n) - \mathbb{E}f(X)| > \epsilon$ along some subsequence $N' \subset \mathbb{N}$. By the relative compactness we may choose a further subsequence $N'' \subset \mathbb{N}$ and a process Y such that $X_n \xrightarrow{d} Y$ along N''. But then $X_n \xrightarrow{fd} Y$ along N'', and since also $X_n \xrightarrow{fd} X$, we must have $X \stackrel{d}{=} Y$ (a consequence of Lemma 9.11). Thus, $X_n \xrightarrow{d} X$ along N'', and so $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ along the same sequence, a contradiction. Thus, we conclude that $X_n \xrightarrow{d} X$. □

Recall that Prohorov's theorem shows that tightness and relative compactness in distribution are equivalent. Also, recall that a set A is *relatively compact* if its closure is compact. Thus, the above result, we need to find convenient criteria for tightness.

The *modulus of continuity* of an arbitrary function $x(\cdot)$ on $[0,1]$ is defined as

$$w(x,h) := \sup_{|s-t|\le h} |x(s) - x(t)|, \qquad h > 0.$$

The function $w(\cdot,h)$ is continuous in $C[0,1]$ and hence a measurable function[102], for each fixed $h > 0$.

We use the classical Arzelá-Ascoli compactness criterion to do this. The Arzelá-Ascoli theorem completely characterizes relative compactness in $C[0,1]$.

Theorem 9.16 (Arzelá-Ascoli)**.** *A set A is relatively compact in $C[0,1]$ if and only if*

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

and

$$\lim_{h\to 0} \sup_{x\in A} w(x,h) = 0. \tag{139}$$

The functions in A are by definition *equicontinuous* at $t_0 \in [0,1]$ if, as $t \to t_0$, $\sup_{x\in A} |x(t) - x(t_0)| \to 0$ (cf. the notion of asymptotic equicontinuity introduced in Definition 1.6); and (139) defines *uniform equicontinuity* (over [0,1]) of the functions in A. We can use it now for characterization of tightness.

Lemma 9.17. *The sequence $\{X_n\}_{n\ge1}$ is tight if and only if the following two conditions hold:*

[102] This follows from the fact that $|w(x,h) - w(y,h)| \le 2\|x-y\|_\infty$.

(i) For every $\eta > 0$, there exists $a \geq 0$ and $n_0 \in \mathbb{N}$ such that

$$\mathbb{P}(|X_n(0)| \geq a) \leq \eta \qquad \text{for all } n \geq n_0. \tag{140}$$

(ii) For each $\epsilon > 0$ and $\eta > 0$, there exist $h \in (0,1)$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$\mathbb{P}(w(X_n, h) \geq \epsilon) \leq \eta \qquad \text{for all } n \geq n_0. \tag{141}$$

Proof. Suppose that $\{X_n\}_{n\geq 1}$ is tight. Given $\eta > 0$, we can find a compact K such that $\mathbb{P}(X_n \in K) > 1-\eta$ for all $n \geq 1$. By the Arzelà-Ascoli theorem, we have $K \subset \{x \in C[0,1] : |x(0)| < a\}$ for large enough a and $K \subset \{x \in C[0,1] : w(x,h) < \epsilon\}$ for small enough h, and so (i) and (ii) hold, with $n_0 = 1$ in each case. Hence the necessity.

Assume then that $\{P_n\}_{n\geq 1}$ satisfies (i) and (ii), with $n_0 = 1$ (without loss of generality[103]), where $X_n \sim P_n$. Given $\eta > 0$, choose a so that, if $B = \{x \in C[0,1] : |x(0)| \leq a\}$, then $P_n(B) \geq 1-\eta$ for all n. Then choose h_k, so that, if $B_k := \{x \in C[0,1] : w(x,h_k) < 1/k\}$, then $P_n(B_k) \geq 1 - \eta/2^k$ for all n. If K is the closure of $A := B \cap (\cap_k B_k)$, then $P_n(K) \geq 1 - 2\eta$ for all n. Since A satisfies the conditions of Theorem 9.16, K is compact. Therefore, $\{P_n\}_{n\geq 1}$ is tight. □

Remark 9.3. *Note that condition (141) can be expressed in a more compact form:*

$$\lim_{h\to 0} \limsup_{n\to\infty} \mathbb{P}(w(X_n, h) > \epsilon) = 0.$$

The following result now gives an equivalent characterization of weak convergence in $C[0,1]$. It is a trivial consequence of what we have established so far.

Theorem 9.18. *Let $X, X_1, X_2, \ldots$ be random elements in $C[0,1]$. Then, $X_n \xrightarrow{d} X$ if and only if $X_n \xrightarrow{fd} X$ and*

$$\lim_{h\to 0} \limsup_{n\to\infty} \mathbb{E}[w(X_n, h) \wedge 1] = 0.$$

Remark 9.4. *Let us fix two metric spaces (K,d) and (S,ρ), where K is compact and S is separable and complete (i.e., S is a Polish space). Everything done in this subsection can be easily extended to the space $C(K,S)$ of continuous functions from K to S, endowed with the uniform metric*

$$\hat{\rho}(x,y) = \sup_{t\in K} \rho(x_t, y_t).$$

[103] Since $C[0,1]$ is separable and complete, a single measure Q is tight (by), and so by the necessity of (i) and (ii), for a given $\eta > 0$ there is an a such that $Q(\{x \in C[0,1] : |x(0)| \geq a\}) \leq \eta$, and for given $\epsilon > 0$ and $\eta > 0$ there is a $h > 0$ such that $Q(\{x \in C[0,1] : w(x,h) > \epsilon\}) \leq \eta$. Therefore, we may ensure that the inequalities in (140) and (141) hold for the finitely many n preceding n_0 by increasing a and decreasing h if necessary. Thus, we may assume that n_0 is always 1.

9.4 Non-measurablilty of the empirical process

Suppose that $U_1, \ldots, U_n$ be i.i.d. Uniform(0,1) defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let G_n be the empirical c.d.f. of the data. Note that G_n (defined on $[0,1]$) is NOT continuous. Thus, the empirical process is not an element of $C[0,1]$. It is standard to consider G_n as an element of $D[0,1]$, the *space of càdlàg functions* (right continuous with left limits) on $[0,1]$. To understand the weak convergence of the empirical process we next study weak convergence on $D[0,1]$.

The following example shows that if $D[0,1]$ is equipped with the Borel σ-field $\mathcal{B}(D[0,1])$ generated by the closed sets under the uniform metric, the empirical d.f. G_n will *not* be a random element in $D[0,1]$, i.e., G_n is not $\mathcal{A}/\mathcal{B}(D[0,1])$-measurable.

Example 9.19. *Consider $n = 1$ so that $G_1(t) = \mathbf{1}_{[0,t]}(U_1)$, $t \in [0,1]$ (visualize the random function G_1 over $[0,1]$; $G_1(t) = \mathbf{1}\{t \geq U_1\}$). Let B_s be the open ball in $D[0,1]$ with center $\mathbf{1}_{[s,1]}$ and radius 1/2, where $s \in [0,1]$. For each subset $A \subset [0,1]$ define*

$$E_A := \cup_{s \in A} B_s.$$

Observe that E_A is an open set as an uncountable union of open sets is also open.

If G_1 were $\mathcal{A}/\mathcal{B}(D[0,1])$-measurable, the set

$$\{\omega \in \Omega :\in G_1(\omega) \in E_A\} = \{\omega \in \Omega : U_1(\omega) \in A\}$$

would belong to $\mathcal{A}$. A probability measure could be defined on the class of all subsets of $[0,1]$ by setting $\mu(A) := \mathbb{P}(U_1 \in A)$. This μ would be an extension of the uniform distribution to all subsets of $[0,1]$. Unfortunately such an extension cannot exist! Thus, we must give up Borel measurability of G_1. The argument can be extended to $n \geq 1$ (see Problem 1 in [Pollard, 1984, Chapter IV]).

The above example shows that the Borel σ-field generated by the uniform metric on $D[0,1]$ contains too many sets.

Exercise (HW4): Show that $D[0,1]$, equipped with the Borel σ-field $\mathcal{B}(D[0,1])$ generated by the uniform metric, is NOT separable. [Hint: We can define $f_x(t) = \mathbf{1}_{[x,1]}(t)$ for each $x \in [0,1]$, then $\|f_x - f_y\|_\infty = 1$ whenever $x \neq y$. In particular, we have an uncountable collection of disjoint open sets given by the balls $B(f_x, 1/2)$, and so the space is not countable.]

Too large a σ-field $\mathcal{T}$ makes it too difficult for a map into T to be a random element. We must also guard against too small a $\mathcal{T}$. Even though the metric space on T has lost the right to have $\mathcal{T}$ equal to the Borel σ-field, it can still demand some degree of compatibility before a fruitful weak convergence theory will result. If $\mathcal{C}_b(T; \mathcal{T})$ contains too few functions, the approximation arguments underlying the continuous mapping theorem will fail. Without that key theorem, weak convergence becomes a barren theory (see [Pollard, 1984, Chapter IV]).

9.5 $D[0,1]$ with the ball σ-field

In the last subsection we saw that the uniform empirical distribution function $G_n \in T := D[0,1]$ is *not measurable* with respect to its Borel σ-field (under the uniform metric). There is a simple alternative to the Borel σ-field that works for most applications in $D[0,1]$, when the limiting distribution is continuous.

For each fixed $t \in [0,1]$, the map $G_n(\cdot, t) : \Omega \to \mathbb{R}$ is a random variable. That is, if π_t denotes the coordinate projection map that takes a function x in $D[0,1]$ onto its value at t, the composition $\pi_t \circ U_n$ is $\mathcal{A}/\mathcal{B}(\mathbb{R})$-measurable.

Let $\mathcal{P}$ be the projection σ-field, i.e., the σ-field generated by the coordinate projection maps. Recall that if $f : T \to \mathbb{R}$, then the *σ-field generated by the function f*, denoted by $\sigma(f)$, is the collection of all inverse images $f^{-1}(B)$ of the sets $B \in \mathcal{B}(\mathbb{R})$, i.e.,

$$\sigma(f) := \{f^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}.$$

Exercise(HW4): A stochastic process X on $(\Omega, \mathcal{A}, \mathbb{P})$ with sample paths in $D[0,1]$, such as an empirical process, is $\mathcal{A}/\mathcal{P}$-measurable provided $\pi_t \circ X$ is $\mathcal{A}/\mathcal{B}(\mathbb{R})$-measurable for each fixed t.

As every cadlag function on $[0,1]$ is bounded (Exercise(HW4)), the uniform distance

$$d(x,y) := \|x - y\| = \sup_{t \in [0,1]} |x(t) - y(t)|$$

defines a metric on $D[0,1]$. Here are a few facts about $\mathcal{P}$ (Exercise(HW4)):

- $\mathcal{P}$ coincides with the σ-field generated by the closed (or open) balls (this is called the *ball σ-field*[104]).
- Every point in $D[0,1]$ is completely regular.

The limit processes for many applications will always concentrate in a separable subset of $D[0,1]$, usually $C[0,1]$, the set of all continuous real valued functions on $[0,1]$.

Exercise(HW4): Show that $C[0,1]$ is a closed, complete, separable, $\mathcal{P}$-measurable subset of $D[0,1]$. Show that for $C[0,1]$, the projection σ-field coincides with the Borel σ-field.

Question: How do we establish convergence in distribution of a sequence $\{X_n\}$ of random elements of $D[0,1]$ to a limit process X?

Definition 9.20. *For each finite subset S of $[0,1]$ write π_S for the projection map from $D[0,1]$ into $\mathbb{R}^S$ that takes an x onto its vector of values $\{x(t) : t \in S\}$.*

Certainly, we need the finite dimensional projections $\{\pi_S X_n\}$ to converge in distribution, as random vectors in $\mathbb{R}^s$, to the finite-dimensional projection of $\pi_S X$, for each finite

[104] For separable spaces, the ball σ-field is the same as the Borel σ-field; see Section 6 of [Billingsley, 1999] for a detailed description and the properties of the ball σ-field.

subset $S \in [0,1]$. The continuity and measurability of π_S and the continuous mapping theorem makes that a necessary condition.

For the sake of brevity, let us now shorten "finite-dimensional distributions" to "fidis" and "finite-dimensional projections" to fidi projections.

Theorem 9.21. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0,1]$ (under its uniform metric and the projection σ-field). Suppose that $\mathbb{P}(X \in C) = 1$ for some separable subset C of $D[0,1]$. The necessary and sufficient conditions for $\{X_n\}$ to converge in distribution to X are:*

(i) $X_n \stackrel{fd}{\to} X$ *(the fidis of X_n converge to the fidis of X, i.e., $\pi_S X_n \stackrel{d}{\to} \pi_S X$ of each finite subset S of $[0,1]$);*

(ii) to each $\epsilon > 0$ and $\delta > 0$ there corresponds a grid $0 = t_0 < t_1 < \ldots < t_m = 1$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right) < \epsilon; \tag{142}$$

where J_i denotes the interval $[t_i, t_{i+1})$, for $i = 0, 1, \ldots, m-1$.

Proof. Suppose that $X_n \stackrel{d}{\to} X$. As the projection map π_S is both continuous and projection-measurable (by definition), the continuous mapping theorem shows that (i) holds.

Let $\epsilon, \delta > 0$ be given. To simplify the proof of (ii) suppose that the separable subset C is $C[0,1]$ (continuity of the sample paths makes the choice of the grid easier). Let $\{s_0, s_1, \ldots\}$ be a countable dense subset of $[0,1]$. We will assume that $s_0 = 0$ and $s_1 = 1$.

For a fixed $x \in C[0,1]$, and given the ordered $k+1$ points $\{s_0, \ldots, s_k\}$ labeled as $0 = t_0 < \ldots < t_k = 1$ in $[0,1]$, define an interpolation $A_k x$ of x as

$$A_k x(t) = x(t_i), \quad \text{for} \quad t_i \le t < t_{i-1},$$

and $A_k x(1) = x(1)$.

For a fixed $x \in C[0,1]$, the distance $\|A_k x - x\|$ converges to zero as k increases, by virtue of the uniform continuity of x. Thus, we can assure the existence of some k for which

$$\mathbb{P}(\|A_k X - X\| \ge \delta) < \epsilon. \tag{143}$$

Choose and fix such a k. As $\|A_k x - x\|$ varies continuously with x (show this!), the set

$$F := \{x \in D[0,1] : \|A_k x - x\| \ge \delta\}$$

is closed. By an application of the Portmanteau theorem (note that this needs a proof, and has not been discussed in class yet, as we are not dealing with the Borel σ-field; see e.g., [Pollard, 1984, Example 17, Chapter IV]; also see [Billingsley, 1999, Theorem 6.3]),

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \le \mathbb{P}(X \in F).$$

The left-hand side bounds the limsup in (142).

Now let us show that (i) and (ii) imply $X_n \xrightarrow{d} X$. Let us retain the assumption that X has continuous sample paths. Choose any bounded, uniformly continuous, projection-measurable, real valued function f on $D[0,1]$. Given $\epsilon > 0$, find $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $\|x - y\| \le \delta$. Write A_T for the approximation map constructed from the grid in (ii) corresponding to this δ and ϵ, i.e.,

$$\limsup_{n\to\infty} \mathbb{P}(\|A_T X_n - X_n\| > \delta) < \epsilon.$$

Without loss of generality we may assume that the A_k of (143) equals A_T.

Now, let us write the composition $f \circ A_T$ as $g \circ \pi_T$, where g is a bounded, continuous function on $D[0,1]$. Thus,

$$\begin{aligned}
& |\mathbb{E}f(X_n) - \mathbb{E}f(X)| \\
\le\ & \mathbb{E}|f(X_n) - f(A_T X_n)| + |\mathbb{E}f(A_T X_n) - \mathbb{E}f(A_T X)| + \mathbb{E}|f(A_T X) - f(X)| \\
\le\ & \epsilon + 2\|f\|\mathbb{P}(\|X_n - A_T X_n\| > \delta) + |\mathbb{E}f(\pi_T X_n) - \mathbb{E}f(\pi_T X)| \\
& \quad + \epsilon + 2\|f\|\mathbb{P}(\|X_n - A_T X_n\| > \delta).
\end{aligned}$$

The middle term in the last line converges to zero as $n \to \infty$, because of the fidi convergence $\pi_T X_n \xrightarrow{d} \pi_T X$. $\square$

We now know the price we pay for wanting to make probability statements about functionals that depend on the whole sample path of a stochastic process: with high probability we need to rule out nasty behavior between the grid points.

Question: How do we control the left-hand side of (142)? It involves a probability of a union of m events, which we may bound by the sum of the probabilities of those events,

$$\sum_{i=0}^{m-1} \mathbb{P}\left(\sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right).$$

Then we can concentrate on what happens in an interval J_i between adjacent grid points. For many stochastic processes, good behavior of the increment $X_n(t_{i+1}) - X_n(T_i)$ forces good behavior for the whole segment of sample path over J_i. Read [Pollard, 1984, Chapter V] for further details and results on how to control the left-hand side of (142).

10 Weak convergence in non-separable metric spaces

We have already seen that the uniform empirical process is not Borel measurable. We have also seen that we can change the Borel σ-field to the ball σ-field and develop a fruitful notion of weak convergence.

Another alternative solution is to use a different metric that will make the empirical process measurable, e.g., we can equip the space $D[0,1]$ with the Skorohod (metric) topology.

Remark 10.1. *($D[0,1]$ and the Skorohod metric) We say that two cadlag functions are close to one another in the Skorohod metric if there is a reparameterization of the time-axis, (a function $[0,1]$ to itself) that is uniformly close to the identity function, and when applied to one of the cadlag functions, brings it close to the other cadlag function. Heuristically, two cadlag functions are close if their large jumps are close to one another and of similar size, and if they are uniformly close elsewhere. See [Billingsley, 1999, Chapter 3] for the study of $D[0,1]$ with the Skorohod metric.*

However, for empirical processes that assume values in very general (and often more complex spaces) such cute generalizations are not readily achievable and the easier way out is to keep the topology simple and tackle the measurability issues. Of course, any generalization of the notion of weak convergence must allow a powerful continuous mapping theorem.

We will now try to develop a more general notion of weak convergence that can handle *non-measurability* of the underlying functions. In this section, the underlying σ-field will always be the Borel σ-field generated by the metric endowed with the space.

Suppose that $\mathbb{D}$ is a metric space with metric $d(\cdot,\cdot)$. Suppose that we have arbitrary maps $X_n : \Omega_n \to \mathbb{D}$, defined on probability spaces $(\Omega_n, \mathcal{A}_n, \mathbb{P}_n)$. Because $\mathbb{E}f(X_n)$ need no longer make sense (where $f : \mathbb{D} \to \mathbb{R}$ is a bounded continuous function), we replace expectations by *outer expectations*.

Definition 10.1 (Outer expectation)**.** *For an arbitrary map $X : \Omega \to \mathbb{D}$, define*

$$\mathbb{E}^* f(X) := \inf \Big\{ \mathbb{E}U \mid U : \Omega \to \mathbb{R} \text{ measurable}, U \geq f(X), \mathbb{E}U \text{ exists} \Big\}. \tag{144}$$

For most purposes outer expectations behave like usual expectations[105].

Definition 10.2. *Then we say that a sequence of arbitrary maps $X_n : \Omega_n \to \mathbb{D}$ converges in distribution*[106] *to a random element X if and only if*

$$\mathbb{E}^* f(X_n) \to \mathbb{E}f(X)$$

[105] Unfortunately, Fubini's theorem is not valid for outer expectations (Note that for any map T it always holds $\mathbb{E}_1^*\mathbb{E}_2^* T \leq \mathbb{E}^* T$). To overcome this problem, it is assumed that $\mathcal{F}$ is P-measurable. We will not define P-measurability formally; see [van der Vaart and Wellner, 1996, Chapter 2.3] for more details.

[106] We note that the weak convergence in the above sense is no longer tied with convergence of probability measures, simply because X_n need not induce Borel probability measures on $\mathbb{D}$.

for every bounded, continuous function $f : \mathbb{D} \to \mathbb{R}$. Here we insist that the limit X be Borel-measurable (and hence has a distribution). Note that although Ω_n may depend on n, we do not let this show up in the notation for $\mathbb{E}^*$ and $\mathbb{P}^*$.

Definition 10.3. *An arbitrary sequence of maps $X_n : \Omega_n \to \mathbb{D}$ converges in probability to X if*

$$\mathbb{P}^*\Big(d(X_n, X) > \epsilon\Big) \to 0 \qquad \text{for all } \epsilon > 0. \tag{145}$$

This will be denoted by $X_n \overset{\mathbb{P}}{\to} X$.

Definition 10.4. *The sequence of maps $X_n : \Omega_n \to \mathbb{D}$ converges almost surely to X if there exists a sequence of (measurable) random variables Δ_n such that*

$$d(X_n, X) \leq \Delta_n \qquad \text{and} \qquad \Delta_n \overset{a.s.}{\to} 0. \tag{146}$$

This will be denoted by $X_n \overset{a.s.^*}{\to} X$.

Remark 10.2. *Even for Borel measurable maps X_n and X, the distance $d(X_n, X)$ need not be a random variable.*

Theorem 10.5 (Portmanteau)**.** *For arbitrary maps $X_n : \Omega_n \to \mathbb{D}$ and every random element X with values in $\mathbb{D}$, the following statements are equivalent:*

(i) $\mathbb{E}^* f(X_n) \to \mathbb{E} f(X)$ *for all real-valued bounded continuous functions f.*

(ii) $\mathbb{E}^* f(X_n) \to \mathbb{E} f(X)$ *for all real-valued bounded Lipschitz functions f.*

(iii) $\liminf_{n\to\infty} \mathbb{P}_*(X_n \in G) \geq \mathbb{P}(X \in G)$ *for every open set G.*

(iv) $\limsup_{n\to\infty} \mathbb{P}^*(X_n \in F) \leq \mathbb{P}(X \in F)$ *for every closed set F.*

(v) $\mathbb{P}^*(X_n \in B) \to \mathbb{P}(X \in B)$ *for all Borel set B with $\mathbb{P}(X \in \partial B) = 0$.*

Proof. See [van der Vaart and Wellner, 1996, Theorem 1.3.4]. □

Theorem 10.6 (Continuous mapping)**.** *Suppose that $H : \mathbb{D} \to S$ (S is a metric space) is a map which is continuous at every $x \in \mathbb{D}_0 \subset \mathbb{D}$. Also, suppose that $X_n : \Omega_n \to \mathbb{D}$ are arbitrary maps and X is a random element with values in $\mathbb{D}_0$ such that $H(X)$ is a random element in S. If $X_n \overset{d}{\to} X$, then $H(X_n) \overset{d}{\to} H(X)$.*

Proof. See [van der Vaart and Wellner, 1996, Theorem 1.3.6]. □

Definition 10.7. *A Borel-measurable random element X into a metric space is tight if for every $\epsilon > 0$ there exists a compact set K such that*

$$\mathbb{P}(X \notin K) < \epsilon.$$

Remark 10.3. *(Separability and tightness) Let $X : \Omega \to \mathbb{D}$ be a random element. Tightness is equivalent to there being a σ-compact set (countable union of compacts) that has probability 1 under X.*

If there is a separable, measurable set with probability 1, then X is called separable. Since a σ-compact set in a metric space is separable, separability is slightly weaker than tightness. The two properties are the same if the metric space is complete[107].

10.1 Bounded stochastic processes

Let us look back at our motivation for studying weak convergence of stochastic processes. Given i.i.d. random elements $X_1, \ldots, X_n \sim P$ taking values in $\mathcal{X}$ and a class of measurable functions $\mathcal{F}$, we want to study the stochastic process $\{\sqrt{n}(\mathbb{P}_n - P)(f) : f \in \mathcal{F}\}$ (i.e., the empirical process). If we assume that

$$\sup_{f\in\mathcal{F}} |f(x) - Pf| < \infty, \qquad \text{for all } x \in \mathcal{X},$$

then the maps from $\mathcal{F}$ to $\mathbb{R}$ defined as

$$f \mapsto f(x) - Pf, \;\; x \in \mathcal{X}$$

are bounded functionals over $\mathcal{F}$, and therefore, so is $f \mapsto (\mathbb{P}_n - P)(f)$. Thus,

$$\mathbb{P}_n - P \in \ell^\infty(\mathcal{F}),$$

where $\ell^\infty(\mathcal{F})$ is the space of bounded real-valued functions on $\mathcal{F}$, a *Banach space*[108] if we equip it with the supremum norm $\|\cdot\|_{\mathcal{F}}$.

Remark 10.4. *As $C[0,1] \subset D[0,1] \subset \ell^\infty[0,1]$, we can consider convergence of a sequence of maps with values in $C[0,1]$ relative to $C[0,1]$, but also relative to $D[0,1]$, or $\ell^\infty[0,1]$. It can be shown that if $\mathbb{D}_0 \subset \mathbb{D}$ be arbitrary metric spaces equipped with the same metric and X (as $\mathbb{D}$) and every X_n take there values in $\mathbb{D}_0$, then $X_n \stackrel{d}{\to} X$ as maps in $\mathbb{D}_0$ if and only if $X_n \to X$ as maps in $\mathbb{D}$.*

A *stochastic process* $X = \{X_t : t \in T\}$ is a collection of random variables $X_t : \Omega \to \mathbb{R}$, indexed by an arbitrary set T and defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For each fixed $\omega \in \Omega$, the map $t \mapsto X_t(\omega)$ is called a *sample path*, and it is helpful to think of X as a random function, whose realizations are the sample paths, rather than a collection of random variables. If every sample path is a *bounded* function, then X can be viewed as a map $X : \Omega \to \ell^\infty(T)$, where $\ell^\infty(T)$ denotes the set of all bounded real-valued functions on T. The space $\ell^\infty(T)$ is equipped with the supremum norm $\|\cdot\|_T$. Unless T is finite, the Banach space $\ell^\infty(T)$ is not separable. Thus the classical theory of convergence in law on

[107] See [Billingsley, 1999, Theorem 1.3] for a proof.

[108] A Banach space is a complete normed vector space.

complete separable metric spaces needs to be extended. Given that the space $\ell^\infty(\cdot)$ arises naturally while studying the empirical process, we will devote this subsection to the study of weak convergence in the space $\ell^\infty(T)$.

It turns out that the theory of weak convergence on bounded stochastic processes extends nicely if the limit laws are assumed to be tight Borel probability measures on $\ell^\infty(T)$. The following result characterizes weak convergence (to a tight limiting probability measure) in $\ell^\infty(T)$: convergence in distribution of a sequence of sample bounded processes is equivalent to weak convergence of the finite dimensional probability laws together with *asymptotic equicontinuity*, a condition that is expressed in terms of probability inequalities. This reduces convergence in distribution in $\ell^\infty(T)$ to maximal inequalities (we have spent most of the earlier part of this course on proving such inequalities). We will often refer to this theorem as the *asymptotic equicontinuity* criterion for convergence in law in $\ell^\infty(T)$. The history of this theorem goes back to Prohorov (for sample continuous processes) and in the present generality it is due to [Hoffmann-Jø rgensen, 1991] with important previous contributions by [Dudley, 1978].

Theorem 10.8. *A sequence of arbitrary maps $X_n : \Omega_n \to \ell^\infty(T)$ converges weakly to a tight random element if and only if both of the following conditions hold:*

(i) The sequence $(X_n(t_1), \ldots, X_n(t_k))$ converges in distribution in $\mathbb{R}^k$ ($k \in \mathbb{N}$) for every finite set of points $t_1, \ldots, t_k \in T$;

(ii) (asymptotic equicontinuity) there exists a semi-metric $d(\cdot,\cdot)$ for which (T, d) is totally bounded and for every $\epsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}^* \left(\sup_{d(s,t)<\delta;\ s,t \in T} |X_n(s) - X_n(t)| > \epsilon \right) = 0. \tag{147}$$

Proof. Let us first show that (i) and (ii) imply that X_n converges weakly to a tight random element. The proof consists of two steps.

Step 1: By assumption (i) and Kolmogorov's consistency theorem[109] we can construct a stochastic process $\{X_t : t \in T\}$ on some probability space such that $(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{d}$

[109] *Kolmogorov's consistency theorem* shows that for any collection of "consistent" finite-dimensional marginal distributions one can construct a stochastic process on some probability space that has these distributions as its marginal distributions.

Theorem 10.9. *Let T denote an arbitrary set (thought of as "time"). For each $k \in \mathbb{N}$ and finite sequence of times $t_1, \ldots t_k \in T$, let $\nu_{t_1,\ldots,t_k}$ denote a probability measure on $\mathbb{R}^k$. Suppose that these measures satisfy two consistency conditions:*

- *for all permutations π of $\{1, \ldots, k\}$ and measurable sets $F_i \subset \mathbb{R}$,*

$$\nu_{\pi(t_1),\ldots,\pi(t_k)}(F_{\pi(1)} \times \ldots \times F_{\pi(k)}) = \nu_{t_1,\ldots,t_k}(F_1 \times \ldots \times F_k);$$

- *for all measurable sets $F_i \subset \mathbb{R}$,*

$$\nu_{t_1,\ldots,t_k}(F_1 \times \ldots \times F_k) = \nu_{t_1,\ldots,t_k,t_{k+1}}(F_1 \times \ldots \times F_k \times \mathbb{R}).$$

$(X(t_1), \ldots, X(t_k))$ for every finite set of points $t_1, \ldots, t_k \in T$. We need to verify that X admits a version that is a tight random element in $\ell^\infty(T)$. Let T_0 be a countable d-dense subset of T, and let $T_k, k = 1, 2, \ldots$ be an increasing sequence of finite subsets of T_0 such that $\cup_{k=1}^\infty T_k = T_0$. By the Portmanteau lemma (see part *(iii)* of Theorem 9.2) on the equivalent conditions of weak convergence in Euclidean spaces we have

$$\begin{aligned}
\mathbb{P}\left(\max_{d(s,t)<\delta;\, s,t\in T_k} |X(s) - X(t)| > \epsilon \right) &\leq \liminf_{n\to\infty} \mathbb{P}\left(\max_{d(s,t)<\delta;\, s,t\in T_k} |X_n(s) - X_n(t)| > \epsilon \right) \\
&\leq \liminf_{n\to\infty} \mathbb{P}\left(\sup_{d(s,t)<\delta;\, s,t\in T_0} |X_n(s) - X_n(t)| > \epsilon \right).
\end{aligned}$$

Taking $k \to \infty$ on the far left side, we conclude that

$$\mathbb{P}\left(\sup_{d(s,t)<\delta;\, s,t\in T_0} |X(s) - X(t)| > \epsilon \right) \leq \liminf_{n\to\infty} \mathbb{P}\left(\sup_{d(s,t)<\delta;\, s,t\in T_0} |X_n(s) - X_n(t)| > \epsilon \right).$$

where we consider the open set $\{x \in \ell^\infty(T) : \sup_{d(s,t)<\delta;\, s,t\in T_k} |x(s) - x(t)| > \epsilon\}$ and by the monotone convergence theorem this remain true if T_k is replaced by T_0.

By the asymptotic equicontinuity condition (147), there exists a sequence $\delta_r > 0$ with $\delta_r \to 0$ as $r \to \infty$ such that

$$\mathbb{P}\left(\sup_{d(s,t)\leq\delta_r,\, s,t\in T_0} |X(s) - X(t)| > 2^{-r} \right) \leq 2^{-r}.$$

These probabilities sum to a finite number over $r \in \mathbb{N}$. Hence by the Borel-Cantelli lemma[110], there exists $r(\omega) < \infty$ a.s. such that for almost all ω,

$$\sup_{d(s,t)\leq\delta_r;\, s,t\in T_0} |X(s;\omega) - X(t;\omega)| \leq 2^{-r}, \qquad \text{for all } r > r(\omega).$$

Hence, $X(t;\omega)$ is a d-uniformly continuous function of t for almost every ω. As T is totally bounded, $X(t;\omega)$ is also bounded. The extension to T by uniform continuity of the restriction of X on T_0 (only the ω set where X is uniformly continuous needs be considered) produces a version of X whose trajectories are all uniformly continuous in (T, d) and, in particular, the law of X admits a tight extension to the Borel σ-algebra of $\ell^\infty(T)$[111].

Then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ *and a stochastic process* $X : T \times \Omega \to \mathbb{R}$ *such that*

$$\mathbb{P}(X_{t_1} \in F_1, \ldots, X_{t_k} \in F_k) = \nu_{t_1,\ldots,t_k}(F_1 \times \ldots \times F_k)$$

for all $t_i \in T$, $k \in \mathbb{N}$ *and measurable set* $F_i \subset \mathbb{R}^n$, *i.e.,* X *has* $\nu_{t_1,\ldots,t_k}$ *as its finite-dimensional distributions relative to times* $t_1, \ldots, t_k$.

[110] Let $\{A_n\}_{n\geq 1}$ be a sequence of some events in some probability space. The Borel-Cantelli lemma states that: If the sum of the probabilities of the events A_n is finite, i.e., $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, then the probability that infinitely many of them occur is 0, i.e., $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$, where $\limsup A_n := \cap_{n=1}^\infty \cup_{k=1}^\infty A_k$ is the set of outcomes that occur infinitely many times within the infinite sequence of events $\{A_n\}$.

[111] We will use the following lemma without proving it.

Step 2: We now prove $X_n \xrightarrow{d} X$ in $\ell^\infty(T)$. First we recall a useful fact[112]: If $f : \ell^\infty(T) \to \mathbb{R}$ is bounded and continuous, and if $K \subset \ell^\infty(T)$ is compact, then for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|u - v\|_T < \delta, \ \ u \in K, v \in \ell^\infty(T) \ \ \Rightarrow \ \ |f(u) - f(v)| < \epsilon. \tag{148}$$

Since (T, d) is totally bounded, for every $\tau > 0$ there exists a finite set of points $t_1, \ldots, t_{N(\tau)}$ which is τ-dense in (T, d) in the sense that $T \subseteq \cup_{i=1}^{N(\tau)} B(t_i, \tau)$, where where $B(t, \tau)$ denotes the open ball of center t and radius τ. Then, for each $t \in T$ we can choose $\pi_\tau : T \to \{t_1, \ldots, t_{N(\tau)}\}$ so that $d(\pi_\tau(t), t) < \tau$. We then define processes $X_{n,\tau}$, $n \in \mathbb{N}$, and X_τ as

$$X_{n,\tau}(t) = X_n(\pi_\tau(t)), \qquad X_\tau(t) = X(\pi_\tau(t)), \quad t \in T. \tag{149}$$

These are approximations of X_n and X that take only a finite number $N(\tau)$ of values. Convergence of the finite dimensional distributions of X_n to those of X implies[113] that

$$X_{n,\tau} \xrightarrow{d} X_\tau \qquad \text{in } \ell^\infty(T). \tag{150}$$

Moreover, the uniform continuity of the sample paths of X implies

$$\lim_{\tau \to 0} \|X - X_\tau\|_T = 0 \quad \text{a.s.} \tag{151}$$

Now let $f : \ell^\infty(T) \to \mathbb{R}$ be a bounded continuous function. We have,

$$\begin{aligned} |\mathbb{E}^* f(X_n) - \mathbb{E} f(X)| &\leq |\mathbb{E}^* f(X_n) - \mathbb{E} f(X_{n,\tau})| + |\mathbb{E} f(X_{n,\tau}) - \mathbb{E} f(X_\tau)| \\ &\qquad + |\mathbb{E} f(X_\tau) - \mathbb{E} f(X)| \\ &\leq I_{n,\tau} + II_{n,\tau} + III_\tau. \end{aligned}$$

We have seen that $\lim_{n\to\infty} II_{n,\tau} = 0$ (by (150)) for each fixed $\tau > 0$ and $\lim_{\tau \to 0} III_\tau = 0$ (as X is uniformly continuous a.s.[114]). Hence it only remains to show that $\lim_{\tau \to 0} \limsup_{n\to\infty} I_{n,\tau} = 0$.

Lemma 10.10. *Let $X(t), t \in T$, be a sample bounded stochastic process. Then the finite dimensional distributions of X are those of a tight Borel probability measure on $\ell^\infty(T)$ if and only if there exists on T a semi-metric d for which (T, d) is totally bounded and such that X has a version with almost all its sample paths uniformly continuous for d.*

[112]Suppose on the contrary that the assertion is false. Then there exist $\epsilon > 0$ and sequences $u_n \in K$ and $v_n \in T$ such that $d(u_n; v_n) \to 0$ and $|f(u_n) - f(v_n)| \geq \epsilon$. Since K is compact, there exists a subsequence $u_{n'}$ of u_n such that $u_{n'}$ has a limit u in K. Then $v_{n'} \to u$ and, by continuity of f, $|f(u_{n'}) - f(v_{n'})| \to |f(u) - f(u)| = 0$, which is a contradiction.

[113]To see this, let $T_i = \{t \in T : t_i = \pi_\tau(t)\}$. Then $\{T_i\}$ forms a partition of T and $\pi_\tau(t) = t_i$ whenever $t \in T_i$ Since the map $\Pi_\tau : (a-1, \ldots, a_{N(\tau)}) \mapsto \sum_{i=1}^{N(\tau)} a_i \mathbf{1}_{T_i}(t)$ is continuous from $\mathbb{R}^{N(\tau)}$ into $\ell^\infty(T)$, the finite dimensional convergence implies that dro any bounded continuous function $H : \ell^\infty(T) \to \mathbb{R}$,

$$\mathbb{E}[H(X_{n,\tau})] = \mathbb{E}[H \circ \Pi_\tau(X_n(t_1), \ldots, X_n(t_{N(\tau)})] \to \mathbb{E}[H \circ \Pi_\tau(X(t_1), \ldots, X(t_{N(\tau)})] = \mathbb{E}[H(X_\tau)],$$

which proves (150).

[114]A more detailed proof goes as follows. Given $\epsilon > 0$, let $K \subset \ell^\infty(T)$ be a compact set such that

Given $\epsilon > 0$ we choose $K \subset \ell^\infty(T)$ to be a compact set such that $\mathbb{P}(X \in K^c) < \epsilon/(6\|f\|_\infty)$. Let $\delta > 0$ be such that (148) holds for K and $\epsilon/6$. Then, we have

$$\begin{aligned} |\mathbb{E}^* f(X_n) - \mathbb{E} f(X_{n,\tau})| &\leq 2\|f\|_\infty \left[\mathbb{P}\Big(X_{n,\tau} \in (K^{\delta/2})^c\Big) + \mathbb{P}\Big(\|X_n - X_{n,\tau}\|_T \geq \delta/2\Big)\right] \\ &\quad + \sup\{|f(u) - f(v)| : u \in K, \|v - v\|_T < \delta\}, \end{aligned} \tag{152}$$

where $K^{\delta/2}$ is the $(\delta/2)$-open neighborhood of the set K under the sup norm, i.e.,

$$K^{\delta/2} := \{v \in \ell^\infty(T) : \inf_{u \in K} \|u - v\|_T < \delta/2\}.$$

The inequality in (152) can be checked as follows: If $X_{n,\tau} \in K^{\delta/2}$ and $\|X_n - X_{n,\tau}\|_T < \delta/2$, then there exists $u \in K$ such that $\|u - X_{n,\tau}\|_T < \delta/2$ and then

$$\|u - X_n\|_T \leq \|u - X_{n,\tau}\|_T + \|X_n - X_{n,\tau}\|_T < \delta.$$

Now the asymptotic equicontinuity hypothesis (see (147); recall the definition of $X_{n,\tau}$ in (149)) implies that there is a $\tau_2 > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}^*\Big(\|X_n - X_{n,\tau}\|_T \geq \delta/2\Big) < \frac{\epsilon}{6\|f\|_\infty}, \qquad \text{for all } \tau < \tau_2.$$

Further, finite-dimensional convergence yields (by the Portmanteau theorem)

$$\limsup_{n\to\infty} \mathbb{P}^*\Big(X_{n,\tau} \in (K^{\delta/2})^c\Big) \leq \mathbb{P}\Big(X_\tau \in (K^{\delta/2})^c\Big) < \frac{\epsilon}{6\|f\|_\infty}.$$

Hence we conclude from (152), that for all $\tau < \min\{\tau_1, \tau_2\}$,

$$\limsup_{n\to\infty} |\mathbb{E}^* f(X_n) - \mathbb{E} f(X_{n,\tau})| < \epsilon,$$

which completes the proof of the part that (i) and (ii) imply the weak convergence of X_n (to tight random element).

Now we show that if X_n converges weakly to a tight random element in $\ell^\infty(T)$, (i) and (ii) should hold. By the continuous mapping theorem it follows that (i) holds. The other implication is a consequence of the "closed set" part of the Portmanteau theorem. First we state a result which we will use (but prove later).

Theorem 10.11. *Suppose that $X \in \ell^\infty(T)$ induces a tight Borel measure. Then there exists a semi-metric ρ on T for which (T, ρ) is totally bounded and such that X has a version with*

$\mathbb{P}(X \in K^c) < \epsilon/(6\|f\|_\infty)$. Let $\delta > 0$ be such that (148) holds for K and $\epsilon/6$. Let $\tau_1 > 0$ be such that $\mathbb{P}(\|X_\tau - X\|_T \geq \delta) < \epsilon/(6\|f\|_\infty)$ for all $\tau < \tau_1$; this can be done by virtue of (151). Then it follows that

$$\begin{aligned} |\mathbb{E} f(X_\tau) - \mathbb{E} f(X)| &\leq 2\|f\|_\infty \mathbb{P}\Big(\{X \in K^c\} \cup \{\|X_\tau - X\|_T \geq \delta\}\Big) + \sup\{|f(u) - f(v)| : u \in K, \|v - v\|_T < \delta\} \\ &\leq 2\|f\|_\infty \Big(\frac{\epsilon}{6\|f\|_\infty} + \frac{\epsilon}{6\|f\|_\infty}\Big) + \frac{\epsilon}{6} = \epsilon. \end{aligned}$$

Hence $\lim_{\tau\to 0} III_\tau = 0$.

almost all sample paths in the space of all uniformly continuous real valued functions from T to $\mathbb{R}$.

Furthermore, if X is zero-mean Gaussian, then this semi-metric can be taken equal to $\rho(s,t) = \sqrt{\mathrm{Var}(X(s) - X(t))}$.

Now, if X_n converges weakly to a tight random element X in $\ell^\infty(T)$, then by Theorem 10.11, there is a semi-metric ρ on T which makes (T, ρ) totally bounded and such that X has (a version with) ρ-uniformly continuous sample paths. Thus for the closed set $F_{\delta,\epsilon}$ defined by

$$F_{\delta,\epsilon} := \Big\{x \in \ell^\infty(T) : \sup_{\rho(s,t)\le\delta} |x(s) - x(t)| \ge \epsilon\Big\},$$

we have (by the Portmanteau theorem)

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\rho(s,t)\le\delta} |X_n(s) - X_n(t)| \ge \epsilon\right) &= \limsup_{n\to\infty} \mathbb{P}^*(X_n \in F_{\delta,\epsilon}) \\ &\le \mathbb{P}(X \in F_{\delta,\epsilon}) = \mathbb{P}\left(\sup_{\rho(s,t)\le\delta} |X(s) - X(t)| \ge \epsilon\right).\end{aligned}$$

Taking limits across the resulting inequality as $\delta \to 0$ yields the asymptotic equicontinuity in view of the ρ-uniform continuity of the sample paths of X.

□

In view of this connection between the partitioning condition (ii), continuity and tightness, we shall sometimes refer to this condition as the condition of *asymptotic tightness* or *asymptotic equicontinuity*.

Remark 10.5. *How do we control the left-hand side of (147)? It involves a probability, which by Markov's inequality can be bounded by*

$$\epsilon^{-1}\mathbb{E}^*\left(\sup_{d(s,t)<\delta} |X_n(s) - X_n(t)|\right).$$

Thus we need good bounds on the expected suprema of the localized fluctuations.

$X \sim P$ is tight is equivalent to the existence of a σ-compact set (a countable union of compacts) that has probability 1 under X. To see this, given $\epsilon = 1/m$, there exists a compact set $E_m \subset \ell^\infty(T)$ such that $P(E_m) > 1 - 1/m$. Take $K_m = \cup_{i=1}^m E_i$ and let $K = \cup_{m=1}^\infty K_m$. Then K_m is an increasing sequence of compacts in $\ell^\infty(T)$. We will show that the semi-metric ρ on T defined by

$$\rho(s,t) = \sum_{m=1}^\infty 2^{-m}(1 \wedge \rho_m(s,t)), \qquad \text{where} \qquad \rho_m(s,t) = \sup_{x\in K_m} |x(s) - x(t)|,$$

where $s, t \in T$ makes (T, ρ) totally bounded. To show this, let $\epsilon > 0$, and choose k so that $\sum_{m=k+1}^\infty 2^{-m} < \epsilon/4$. By the compactness of K_k, there exists $x_1, \ldots, x_r$, a finite subset of

K_k, such that $K_k \subset \cup_{i=1}^r B(x_i; \epsilon/4)$ (here $B(x;\epsilon) := \{y \in \ell^\infty(T) : \|x-y\|_T < \epsilon\}$ is the ball of radius ϵ around x), i.e., for each $x \in K_k = \cup_{m=1}^k K_m$ there exists $i \in \{1,\dots,r\}$ such that

$$\|x - x_i\|_T \leq \frac{\epsilon}{4}. \tag{153}$$

Also, as K_k is compact, K_k is a bounded set. Thus, the subset $A \subset \mathbb{R}^r$ defined by $\{(x_1(t),\dots,x_r(t)) : t \in T\}$ is bounded. Therefore, A is totally bounded and hence there exists a finite set $T_\epsilon := \{t_j : 1 \leq j \leq N\}$ such that, for every $t \in T$, there is a $j \leq N$ for which

$$\max_{i=1,\dots,r} |x_i(t) - x(t_j)| \leq \epsilon/4. \tag{154}$$

Next we will show that T_ϵ is ϵ-dense in T for the semi-metric ρ. Let $t \in T$. For any $m \leq k$, we have

$$\rho_m(t, t_j) = \sup_{x \in K_m} |x(t) - x(t_j)| \leq \frac{3\epsilon}{4}. \tag{155}$$

Note that (155) follows as for any $x \in K_m$, there exists $i \in \{1,\dots,r\}$ such that $\|x - x_i\|_T \leq \epsilon/4$ (by (153)) and thus,

$$|x(t) - x(t_j)| \leq |x(t) - x_i(t)| + |x_i(t) - x_i(t_j)| + |x_i(t_j) - x(t_j)| \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{3\epsilon}{4},$$

where we have used (154). Hence,

$$\rho(t, t_j) \leq \sum_{m=1}^k 2^{-m} \rho_m(t, t_j) + \sum_{m=k+1}^\infty 2^{-m} \leq \frac{3\epsilon}{4} \sum_{m=1}^k 2^{-m} + \frac{\epsilon}{4} \leq \epsilon.$$

Thus, we have proved that (T, ρ) is totally bounded.

Furthermore, the functions $x \in K$ are uniformly ρ-continuous, since, if $x \in K_m$, then $|x(s) - x(t)| \leq \rho_m(s,t) \leq 2^m \rho(s,t)$ for all $s, t \in T$ with $\rho(s,t) \leq 1$ (which is always the case). Since $P(K) = 1$, the identity function of $(\ell^\infty(T), \mathcal{B}, P)$ yields a version of X with almost all of its sample paths in K, hence in $UC(T, \rho)$.

Now let ρ_2 be the standard deviation semi-metric. Since every uniformly ρ-continuous function has a unique continuous extension to the ρ-completion of T, which is compact, it is no loss of generality to assume that T is ρ-compact. Let us also assume that *every* sample path of X is ρ-continuous.

An arbitrary sequence $\{t_n\}_{n\geq 1}$ in T has a ρ-converging subsequence $t_{n'} \to t$. By the ρ-continuity of the sample paths, $X(t_{n'}) \to X(t)$ almost surely. Since every $X(t)$ is Gaussian, this implies convergence of means and variances, whence $\rho_2(t_{n'}, t)^2 = \mathbb{E}(X(t_{n'}) - X(t))^2 \to 0$ by a convergence lemma. Thus $t_{n'} \to t$ also for ρ_2, and hence T is ρ_2-compact.

Suppose that a sample path $t \mapsto X(t,\omega)$ is not ρ_2-continuous. Then there exists an $\epsilon > 0$ and a $t \in T$ such that $\rho_2(t_n, t) \to 0$, but $|X(t_n,\omega) - X(t,\omega)| \geq \epsilon$ for every n. By the ρ-compactness and continuity, there exists a subsequence such that $\rho(t_{n'}, s) \to 0$ and $X(t_{n'},\omega) \to X(s,\omega)$ for some $s \in T$. By the argument of the preceding paragraph,

$\rho_2(t_{n'}, s) \to 0$, so that $\rho_2(s,t) = 0$ and $|X(s,\omega) - X(t,\omega)| \geq \epsilon$. Conclude that the path $t \mapsto X(t,\omega)$ can only fail to be ρ_2-continuous for ω for which there exist $s, t \in T$ with $\rho_2(s,t) = 0$, but $X(s,\omega) \neq X(t,\omega)$. Let N be the set of ω for which there do exist such s and t. Take a countable, ρ-dense subset A of $\{(s,t) \in T \times T : \rho_2(s,t) = 0\}$. Since $t \mapsto X(t,\omega)$ is ρ-continuous, N is also the set of all ω such that there exist $(s,t) \in A$ with $X(s,\omega) \neq X(t,\omega)$. From the definition of ρ_2, it is clear that for every fixed (s,t), the set of ω such that $X(s,\omega) \neq X(t,\omega)$ is a null set. Conclude that N is a null set. Hence, almost all paths of X are ρ_2-continuous. $\square$

Remark 10.6. *In the course of the proof of the preceding theorem we constructed a semi-metric ρ such that the weak limit X has uniformly ρ-continuous sample paths, and such that (T, ρ) is totally bounded. This is surprising: even though we are discussing stochastic processes with values in the very large space $\ell^\infty(T)$, the limit is concentrated on a much smaller space of continuous functions. Actually this is a consequence of imposing the condition (ii) and insisting that the limit X be a tight random element.*

10.2 Spaces of locally bounded functions

Let $T_1 \subset T_2 \subset \ldots$ be arbitrary sets and $T = \cup_{i=1}^\infty T_i$. Think of $T_i = [-i, i]$ in which case $T = \mathbb{R}$. The space $\ell^\infty(T_1, T_2, \ldots)$ is defined as the set of all functions $z : T \to \mathbb{R}$ that are uniformly bounded on every T_i (but not necessarily on T).

Exercise (HW4): Show that $\ell^\infty(T_1, T_2, \ldots)$ is a complete metric space with respect to the metric

$$d(z_1, z_2) := \sum_{i=1}^{\infty} (\|z_1 - z_2\|_{T_i} \wedge 1)\, 2^{-i}.$$

Exercise (HW4): Show that a sequence converges in this metric if it converges uniformly on each T_i.

In case $T_i := [-i, i]^d \subset \mathbb{R}^d$ (for $d \geq 1$), the metric d induces the topology of *uniform convergence on compacta*.

The space $\ell^\infty(T_1, T_2, \ldots)$ is of interest in applications, but its weak convergence theory is uneventful. Weak convergence of a sequence is equivalent to (weak) convergence in each of the restrictions T_i's.

Theorem 10.12. *Let $X_n : \Omega_n \to \ell^\infty(T_1, T_2, \ldots)$ be arbitrary maps, $n \geq 1$. Then the sequence $\{X_n\}_{n \geq 1}$ converges weakly to a tight limit if and only if for every $i \in \mathbb{N}$, $X_{n|T_i} : \Omega_n \to \ell^\infty(T_i)$ converges weakly to a tight limit.*

Proof. See [van der Vaart and Wellner, 1996, Theorem 1.6.1]. $\square$

11 Donsker classes of functions

Suppose that $X_1, \ldots, X_n$ are i.i.d. random elements taking values in a set $\mathcal{X}$ having distribution P and let $\mathbb{G}_n$ denote the corresponding empirical process (i.e., $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$) indexed by a class $\mathcal{F}$ of real-valued measurable functions $\mathcal{F}$.

Definition 11.1. *A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is Donsker if the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ converges in distribution in the space $\ell^\infty(\mathcal{F})$ to a tight random element.*

The Donsker property depends on the law P of the observations; to stress this we also say "P-Donsker". The definition implicitly assumes that the empirical process can be viewed as a map into $\ell^\infty(\mathcal{F})$, i.e., that the sample paths $f \mapsto \mathbb{G}_n f$ are bounded. By the multivariate CLT, for any finite set of measurable functions f_i with $P f_i^2 < \infty$,

$$(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k) \xrightarrow{d} (\mathbb{G}_P f_1, \ldots, \mathbb{G}_P f_k),$$

where the vector on the right-hand side possesses a multivariate normal distribution with mean zero and covariances given by

$$\mathbb{E}[\mathbb{G}_P f \mathbb{G}_P g] = P(fg) - (Pf)(Pg).$$

Remark 11.1. *A stochastic process $\mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$ is called Gaussian if for every $(f_1, \ldots, f_k)$, $f_i \in \mathcal{F}$, $k \in \mathbb{N}$, the random vector $(\mathbb{G}_P(f_1), \ldots, \mathbb{G}_P(f_k))$ is a multivariate normal vector.*

As an immediate consequence of Theorem 10.8, we have the following result.

Theorem 11.2. *Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ such that $P[f^2] < \infty$, for every $f \in \mathcal{F}$, and*

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty, \qquad \text{for all } x \in \mathcal{X}.$$

Then the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly to a tight random element (i.e., $\mathcal{F}$ is P-Donsker) if and only if there exists a semi-metric $d(\cdot, \cdot)$ on $\mathcal{F}$ such that $(\mathcal{F}, d)$ is totally bounded and

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}^* \left(\sup_{d(f,g) \le \delta;\; f,g \in \mathcal{F}} |\mathbb{G}_n(f - g)| > \epsilon \right) = 0, \qquad \text{for every } \epsilon > 0. \tag{156}$$

A typical distance d is $d(f, g) = \|f - g\|_{L_2(P)}$, but this is not the only one.

11.1 Donsker classes under bracketing condition

For most function classes of interest, the bracketing numbers $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$ grow to infinity as $\epsilon \downarrow 0$. A sufficient condition for a class to be Donsker is that they do not grow too fast. The speed can be measured in terms of the *bracketing integral*. Recall that the bracketing entropy integral is defined as

$$J_{[\,]}(\delta, \mathcal{F}, L_2(P)) := \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F} \cup \{0\}, L_2(P))}\, d\epsilon.$$

If this integral is finite-valued, then the class $\mathcal{F}$ is P-Donsker. As the integrand is a decreasing function of ϵ the convergence of the integral depends only on the size of the bracketing numbers for $\epsilon \downarrow 0$. Because $\int_0^1 \epsilon^{-r} d\epsilon$ converges for $r < 1$ and diverges for $r \geq 1$, the integral condition roughly requires that the entropies grow of slower order than $1/\epsilon^2$.

Theorem 11.3 (Donsker theorem)**.** *Suppose that $\mathcal{F}$ is a class of measurable functions with square-integrable (measurable) envelope F and such that $J_{[\,]}(1, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is P-Donsker.*

Proof. As $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$ is finite for every $\epsilon > 0$, we know that $(\mathcal{F}, d)$ is totally bounded, where $d(f, g) = \|f - g\|_{L_2(P)}$. Let $\mathcal{G}$ be the collection of all differences $f - g$ when f and g range over $\mathcal{F}$. With a given set of ϵ-brackets $\{[l_i, u_i]\}_{i=1}^N$ over $\mathcal{F}$ we can construct 2ϵ-brackets over $\mathcal{G}$ by taking differences $[l_i - u_j, u_i - l_j]$ of upper and lower bounds. Therefore, the bracketing numbers $N_{[\,]}(\epsilon, \mathcal{G}, L_2(P))$ are bounded by the squares of the bracketing numbers $N_{[\,]}(\epsilon/2, \mathcal{F}, L_2(P))$. Taking a logarithm turns the square into a multiplicative factor 2 and hence the entropy integrals of $\mathcal{F}$ and $\mathcal{G}$ are proportional. The function $G = 2F$ is an envelope for the class $\mathcal{G}$.

Let $\mathcal{G}_\delta := \{f - g : f, g \in \mathcal{F}, \|f - g\|_{L_2(P)} \leq \delta\}$. Hence, by a maximal inequality[115], there exists a finite number $a(\delta) = \delta/\sqrt{\log N_{[\,]}(\delta, \mathcal{G}_\delta, L_2(P))}$ such that

$$\begin{aligned}\mathbb{E}^*\Big[\sup_{f,g \in \mathcal{F}; \|f-g\| \leq \delta} |\mathbb{G}_n(f - g)|\Big] &\lesssim J_{[\,]}(\delta, \mathcal{G}_\delta, L_2(P)) + \sqrt{n} P[G \mathbf{1}\{G > a(\delta)\sqrt{n}\}]. \\ &\leq J_{[\,]}(\delta, \mathcal{G}, L_2(P)) + \sqrt{n} P[G \mathbf{1}\{G > a(\delta)\sqrt{n}\}]. \qquad (157)\end{aligned}$$

The second term on the right is bounded by $a(\delta)^{-1} P[G^2 \mathbf{1}\{G > a(\delta)\sqrt{n}\}]$ and hence converges to 0 as $n \to \infty$ for every δ. The integral converges to zero as $\delta \to 0$. The theorem now follows from the asymptotic equi-continuity condition (see Theorem 11.2), in view of Markov's inequality. □

Example 11.5 (Classical Donsker's theorem)**.** *When $\mathcal{F}$ is equal to the collection of all indicator functions of the form $f_t = \mathbf{1}_{(-\infty, t]}$, with t ranging over $\mathbb{R}$, then the empirical*

[115] Here is a maximal inequality that uses bracketing entropy (see [van der Vaart, 1998, Lemma 19.34] for a proof):

Theorem 11.4. *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $Pf^2 < \delta^2$, for every f, we have, with $a(\delta) = \delta/\sqrt{\log N_{[\,]}(\delta, \mathcal{F}, L_2(P))}$, and F an envelope function,*

$$\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}} = \mathbb{E}^*\Big[\sup_{f \in \mathcal{F}} |\mathbb{G}_n f|\Big] \lesssim J_{[\,]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n} P^*[F \mathbf{1}\{F > \sqrt{n} a(\delta)\}].$$

process $\mathbb{G}_n f_t$ is the classical empirical process $\sqrt{n}(\mathbb{F}_n(t) - F(t))$ (here $X_1, \ldots, X_n$ are i.i.d. P with c.d.f. F).

We saw previously that $N_{[\,]}(\sqrt{\epsilon}, \mathcal{F}, L_2(P)) \leq 2/\epsilon$, whence the bracketing numbers are of the polynomial order $1/\epsilon^2$. This means that this class of functions is very small, because a function of the type $\log(1/\epsilon)$ satisfies the entropy condition of Theorem 5.2 easily.

11.2 Donsker classes with uniform covering numbers

The following theorem shows that the bracketing numbers in the preceding Donsker theorem can be replaced by the *uniform covering numbers*[116]

$$\sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)).$$

Here the supremum is taken over all probability measures Q for which $\|F\|_{Q,2}^2 = Q[F^2] > 0$. Recall that the *uniform entropy integral* is defined as

$$J(\delta, \mathcal{F}, F) = \int_0^\delta \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon. \tag{158}$$

Theorem 11.6. *(Donsker theorem). Let $\mathcal{F}$ be a pointwise-measurable*[117] *class of measurable functions with (measurable) envelope F such that $P[F^2] < \infty$. If $J(1, \mathcal{F}, F) < \infty$ then $\mathcal{F}$ is P-Donsker.*

Proof. We first show that $(\mathcal{F}, d)$ is totally bounded, where $d(f, g) = \|f - g\|_{L_2(P)}$. The finiteness of the uniform entropy integral implies the finiteness of its integrand: $\sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) < \infty$, for every $\epsilon > 0$, where the supremum is taken over all finitely discrete probability measures Q with $Q[F^2] > 0$. We claim that this implies that $\mathcal{F}$ is totally bounded in $L_2(P)$. Let $\epsilon > 0$ and suppose that $f_1, \ldots, f_N$ are functions in $\mathcal{F}$ such that $P[(f_i - f_j)^2] > \epsilon^2 P[F^2]$, for every $i \neq j$. By the law of large numbers $\mathbb{P}_n[(f_i - f_j)^2] \to P[(f_i - f_j)^2]$ for any i, j and $\mathbb{P}_n[F^2] \to PF^2$, almost surely, as $n \to \infty$. It follows that there exists some n and realization P_n of $\mathbb{P}_n$ such that $P_n[(f_i - f_j)^2] > \epsilon^2 P[F^2]$, for every $i \neq j$, and $0 < P_n[F^2] < 2P[F^2]$. Consequently $P_n[(f_i - f_j)^2] > \epsilon^2 P_n[F^2]/2$, and hence $N \leq D(\epsilon \|F\|_{P_n,2}/\sqrt{2}, \mathcal{F}, L_2(P_n))$ (recall the notion of packing number; see Definition 2.5). Because P_n is finitely discrete, the right side is bounded by the supremum over Q considered previously and hence bounded in n. In view of the definition of N this shows that $D(\epsilon \|F\|_{P,2}, \mathcal{F}, L_2(P))$ is finite for every $\epsilon > 0$.

[116] The uniform covering numbers are relative to a given envelope function F. This is fortunate, because the covering numbers under different measures Q typically are more stable if standardized by the norm $\|F\|_{Q,2}$ of the envelope function. In comparison, in the case of bracketing numbers we consider a single distribution P.

[117] The condition that the class $\mathcal{F}$ is pointwise-measurable (or "suitably measurable") is satisfied in most examples but cannot be omitted. It suffices that there exists a countable collection $\mathcal{G}$ of functions such that each f is the pointwise limit of a sequence g_m in $\mathcal{G}$; see [van der Vaart and Wellner, 1996, Chapter 2.3].

To verify the asymptotic equi-continuity condition, it is enough to show that

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{E}[\sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{G}_\delta}] = 0,$$

where $\mathcal{G}_\delta := \{f - g : f, g \in \mathcal{F}, \|f - g\|_{L_2(P)} \le \delta\}$. The class $\mathcal{G}_\delta$ has envelope $2F$, and since $\mathcal{G}_\delta \subset \{f - g : f, g \in \mathcal{F}\}$, we have

$$\sup_Q N(\epsilon\|2F\|_{Q,2}, \mathcal{G}_\delta, \|\cdot\|_{Q,2}) \le \sup_Q N^2(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}),$$

which leads to $J(\epsilon, \mathcal{G}_\delta, 2F) \le CJ(\epsilon, \mathcal{F}, F)$ for all $\epsilon > 0$. Hence by the maximal inequality in Theorem 4.10 (with $\sigma = \delta$, envelope $2F$) and $\delta' = \delta/(2\|F\|_{P,2})$,

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{G}_\delta}] \le C\left\{J(\delta', \mathcal{F}, F)\|F\|_{P,2} + \frac{B_n J^2(\delta', \mathcal{F}, F)}{\delta'^2\sqrt{n}}\right\},$$

where $B_n = 2\sqrt{\mathbb{E}[\max_{1\le i\le n} F^2(X_i)]}$. As $F \in L_2(P)$, we have[118] $B_n = o(\sqrt{n})$. Given $\eta > 0$, by choosing δ small, we can make sure that $J(\delta', \mathcal{F}, F) < \eta$ and for large n, $B_n J^2(\delta', \mathcal{F}, F)/(\delta'^2\sqrt{n}) < \eta$. Thus,

$$\limsup_{n\to\infty}\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{G}_\delta}] \le (C\|F\|_{P,2} + 1)\eta,$$

so that the desired conclusion follows.

Alternate proof: The entropy integral of the class $\mathcal{F} - \mathcal{F}$ (with envelope $2F$) is bounded by a multiple of $J(\delta, \mathcal{F}, F)$. Application of Theorem 4.7 followed by the Cauchy-Schwarz inequality yields, with $\theta_n^2 := \sup_{f\in\mathcal{G}_\delta}\mathbb{P}_n[f^2]/\|F\|_n^2$,

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{G}_\delta}\Big] \lesssim \mathbb{E}\Big[J(\theta_n, \mathcal{G}_\delta, 2F)\|F\|_n\Big] \lesssim \mathbb{E}^*[J^2(\theta_n, \mathcal{G}_\delta, 2F)]^{1/2}\|F\|_{P,2}.$$

If $\theta_n \xrightarrow{\mathbb{P}} 0$, then the right side converges to zero, in view of the dominated convergence theorem.

Without loss of generality, assume that $F \ge 1$, so that $\theta_n^2 \le \sup_{f\in\mathcal{G}_\delta}\mathbb{P}_n[f^2]$ (otherwise replace F by $F \vee 1$; this decreases the entropy integral). As $\sup_{f\in\mathcal{G}_\delta} P[f^2] \to 0$, as $\delta \to 0$, the desired conclusion follows if $\|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{G}_\delta}$ converges in probability to zero. This is certainly the case if the class $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$ is Glivenko-Cantelli. It can be shown $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$, relative to the envelope $(2F)^2$, has bounded uniform entropy integral. Thus the class is Glivenko-Cantelli. $\square$

11.3 Donsker theorem for classes changing with sample size

The Donsker theorem we just stated involves a fixed class of functions $\mathcal{F}$ not depending on n. As will become clear in the next example, it is sometime useful to have similar results

[118]This follows from the following simple fact: For i.i.d. random variables $\xi_1, \xi_2, \ldots$ the following three statements are equivalent: (1) $\mathbb{E}[|\xi_1|] < \infty$; (2) $\max_{1\le i\le n}|\xi_i|/n \to 0$ almost surely; (3) $\mathbb{E}[\max_{1\le i\le n}|\xi_i|] = o(n)$.

for classes of functions $\mathcal{F}_n$ which depend on the sample size n. Suppose that

$$\mathcal{F}_n := \{f_{n,t} : t \in T\}$$

is a sequence of function classes (indexed by T) where each $f_{n,t}$ is a measurable function from $\mathcal{X}$ to $\mathbb{R}$. We want to treat the weak convergence of the stochastic processes $\mathcal{Z}_n$ defined as

$$\mathcal{Z}_n(t) = \mathbb{G}_n f_{n,t}, \qquad t \in T \tag{159}$$

as elements of $\ell^\infty(T)$. We know that weak convergence in $\ell^\infty(T)$ is equivalent to marginal convergence and asymptotic equi-continuity. The marginal convergence to a Gaussian process follows under the conditions of the Lindeberg-Feller CLT[119]. Sufficient conditions for equi-continuity can be given in terms of the entropies of the classes $\mathcal{F}_n$.

We will assume that there is a semi-metric ρ for the index set T for which (T, ρ) is totally bounded, and such that

$$\sup_{\rho(s,t)<\delta_n} P(f_{n,s} - f_{n,t})^2 \to 0 \qquad \text{for every} \quad \delta_n \to 0. \tag{160}$$

Suppose further that the classes $\mathcal{F}_n$ have envelope functions F_n satisfying the Lindeberg condition

$$P[F_n^2] = O(1), \qquad \text{and} \qquad P[F_n^2 \mathbf{1}_{\{F_n > \epsilon\sqrt{n}\}}] \to 0 \;\text{ for every } \epsilon > 0. \tag{161}$$

Theorem 11.7. *Suppose that $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ is a class of measurable function indexed by (T, ρ) which is totally bounded. Suppose that (160) and (161) hold. If $J_{[\,]}(\delta_n, \mathcal{F}_n, L_2(P)) \to 0$ for every $\delta_n \to 0$, or $J(\delta_n, \mathcal{F}_n, F_n) \to 0$ for every $\delta_n \to 0$ and all the classes $\mathcal{F}_n$ are P-measurable[120], then the processes $\{\mathcal{Z}_n(t) : t \in T\}$ defined by (159) converge weakly to a tight Gaussian process $\mathcal{Z}$ provided that the sequence of covariance functions*

$$K_n(s,t) = P(f_{n,s}, f_{n,t}) - P(f_{n,s})P(f_{n,t})$$

converges pointwise on $T \times T$. If $K(s,t)$, $s,t \in T$, denotes the limit of the covariance functions, then it is a covariance function and the limit process $\mathcal{Z}$ is a mean zero Gaussian process with covariance function K.

[119]**Lindeberg-Feller CLT:** For each $n \in \mathbb{N}$, suppose that $W_{n,1}, \ldots, W_{n,k_n}$ are independent random vectors with finite variances such that

$$\sum_{i=1}^{k_n} \mathbb{E}[\|W_{n,i}\|^2 \mathbf{1}\{\|W_{n,i}\| > \epsilon\}] \to 0, \qquad \text{for every } \epsilon > 0, \qquad \text{and} \qquad \sum_{i=1}^{k_n} \mathrm{Cov}(W_{n,i}) \to \Sigma.$$

Then the sequence $\sum_{i=1}^{k_n}(W_{n,i} - \mathbb{E}W_{n,i})$ converges in distribution to a normal $N(0, \Sigma)$ distribution.

[120]We will not define P-measurability formally; see [van der Vaart and Wellner, 1996, Chapter 2.3] for more details. However if $\mathcal{F}$ is point-wise measurable then $\mathcal{F}$ is P-measurable for every P.

Proof. We only give the proof using the bracketing entropy integral condition. For each $\delta > 0$, applying Theorem 4.12 and using a similar idea as in (157) we obtain the bound

$$\mathbb{E}^*\Big[\sup_{s,t\in T;\, P(f_{n,s}-f_{n,t})^2<\delta^2} |\mathbb{G}_n(f_{n,s}-f_{n,t})|\Big] \lesssim J_{[\,]}(\delta, \mathcal{G}_n, L_2(P)) + a_n(\delta)^{-1} P[G_n^2 \mathbf{1}\{G_n > a_n(\delta)\sqrt{n}\}],$$

where $\mathcal{G}_n = \mathcal{F}_n - \mathcal{F}_n$, $G_n = 2F_n$ and $a_n(\delta) = \delta/\sqrt{\log N_{[\,]}(\delta, \mathcal{G}_n, L_2(P))}$. Because $J_{[\,]}(\delta_n, \mathcal{F}_n, L_2(P)) \to 0$ for every $\delta_n \to 0$, we must have that $J_{[\,]}(\delta, \mathcal{F}_n, L_2(P)) = O(1)$ for every $\delta > 0$ and hence $a_n(\delta)$ is bounded away from 0. Then the second term in the preceding display converges to zero for every fixed $\delta > 0$, by the Lindeberg condition. The first term can be made arbitrarily small as $n \to \infty$ by choosing δ small, by assumption. □

Example 11.8 (The Grenander estimator). *Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $[0,\infty)$ with a non-increasing density function f and c.d.f. F (which is known to be concave). We want to estimate the unknown density f under the restriction that f is non-increasing.*

Grenander [Grenander, 1956] showed that we can find a nonparametric maximum likelihood estimator (NPMLE) $\hat{f}_n$ in this problem, i.e., we can maximize the likelihood $\prod_{i=1}^n g(X_i)$ over all non-increasing densities g on $[0,\infty)$.

Let $\mathbb{F}_n$ is the empirical distribution function of the data. It can be shown that $\hat{f}_n$ is unique and that $\hat{f}_n$ is the left derivative of the least concave majorant (LCM) of $\mathbb{F}_n$, i.e.,

$$\hat{f}_n = LCM'[\mathbb{F}_n];$$

see [Robertson et al., 1988, Chapter 7.2]. Also, $\hat{f}_n$ can be computed easily using the pool adjacent violators algorithm.

Suppose that $x_0 \in (0,\infty)$ is an interior point in the support of P. Does $\hat{f}_n(x_0) \to f(x_0)$? Indeed, this holds. In fact, it can be shown that $n^{1/3}(\hat{f}_n(x_0) - f(x_0)) = O_p(1)$.

Let us find the limiting distribution of $\Delta_n := n^{1/3}(\hat{f}_n(x_0) - f(x_0))$. We will show that if $f'(x_0) < 0$, then

$$\Delta_n = n^{1/3}(\hat{f}_n(x_0) - f(x_0)) \xrightarrow{d} LCM'[\mathcal{Z}](0),$$

where $\mathcal{Z}(s) = \sqrt{f(x_0)}\mathbb{W}(s) + s^2 f'(x_0)/2$, $\mathbb{W}$ is a two-sided standard Brownian motion starting at 0.

Let us define the stochastic process

$$\mathcal{Z}_n(t) := n^{2/3}\left(\mathbb{F}_n(x_0 + tn^{-1/3}) - \mathbb{F}_n(x_0) - f(x_0)tn^{-1/3}\right), \qquad \text{for } t \geq -x_0 n^{-1/3}.$$

Observe that $\Delta_n = LCM'[\mathcal{Z}_n](0)$. Here we have used the fact that for any function $m : \mathbb{R} \to \mathbb{R}$ and an affine function $x \mapsto a(x) := \beta_0 + \beta_1 x$, $LCM[m+a] = LCM[m] + a$.

The idea is to show that

$$\mathcal{Z}_n \xrightarrow{d} \mathcal{Z} \qquad \text{in } \ell^\infty([-K,K]), \qquad \text{for any } K > 0, \tag{162}$$

and then apply the continuous mapping principle to deduce $\Delta_n \stackrel{d}{\to} LCM[\mathcal{Z}]'(0)$.

Actually, a rigorous proof of the convergence of Δ_n *involves a little more than an application of a continuous mapping theorem. The convergence* $\mathcal{Z}_n \stackrel{d}{\to} \mathcal{Z}$ *is only under the metric of uniform convergence on compacta. A concave majoring near the origin might be determined by values of the process a long way from the origin; thus the convergence* $\mathcal{Z}_n \stackrel{d}{\to} \mathcal{Z}$ *by itself does not imply the convergence of* $LCM'[\mathcal{Z}_n](0) \stackrel{d}{\to} LCM[\mathcal{Z}]'(0)$. *However, we will not address this issue for the time being. An interested reader can see [Kim and Pollard, 1990, Assertion, Page 217] for a rigorous treatment of this issue. We will try to show that (162) holds.*

Consider the class of functions

$$g_\theta(x) = \mathbf{1}_{(-\infty,x_0+\theta]}(x) - \mathbf{1}_{(-\infty,x_0]}(x) - f(x_0)\theta,$$

where $\theta \in \mathbb{R}$. *Note that*

$$\begin{aligned} \mathcal{Z}_n(t) &= n^{2/3}\mathbb{P}_n[g_{tn^{-1/3}}(X)] \\ &= n^{2/3}(\mathbb{P}_n - P)[g_{tn^{-1/3}}(X)] + n^{2/3}P[g_{tn^{-1/3}}(X)]. \end{aligned}$$

Let $\mathcal{F}_n := \{f_{n,t} := n^{1/6}g_{tn^{-1/3}} : t \in \mathbb{R}\}$. *It can be shown, appealing to Theorem 11.7, that* $n^{2/3}(\mathbb{P}_n - P)[g_{tn^{-1/3}}(X)] = \mathbb{G}_n[f_{n,t}] \stackrel{d}{\to} \sqrt{f(x_0)}\mathbb{W}(t)$ *in* $\ell^\infty[-K,K]$, *for every* $K > 0$. *Further, using a Taylor series expansion, we can see that*

$$n^{2/3}P[g_{tn^{-1/3}}(X)] = n^{2/3}\left[F(x_0 + tn^{-1/3}) - F(x_0) - n^{-1/3}tf(x_0)\right] \to \frac{t^2}{2}f'(x_0),$$

uniformly on compacta. Combining these two facts we see that (162) holds.

12 Limiting distribution of M-estimators

Let $X_1, \ldots, X_n$ be i.i.d. P observations taking values in a space $\mathcal{X}$. Let Θ denote a parameter space (assumed to be a metric space with metric $d(\cdot,\cdot)$) and, for each $\theta \in \Theta$, let m_θ denote a real-valued function on $\mathcal{X}$. Consider the map

$$\theta \mapsto \mathbb{M}_n(\theta) := \mathbb{P}_n[m_\theta(X)] \equiv \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$$

and let $\hat{\theta}_n$ denote the *maximizer* of $\mathbb{M}_n(\theta)$ over $\theta \in \Theta$, i.e.,

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \mathbb{M}_n(\theta).$$

Such a quantity $\hat{\theta}_n$ is called an *M-estimator*. We study the (limiting) distribution of M-estimators (properly standardized) in this section.

Their statistical properties of $\hat{\theta}_n$ depend crucially on the behavior of the criterion function $\mathbb{M}_n(\theta)$ as $n \to \infty$. For example, we may ask: is $\hat{\theta}_n$ converging to some $\theta_0 \in \Theta$, as $n \to \infty$? A natural way to tackle the question is as follows: We expect that for each $\theta \in \Theta$, $\mathbb{M}_n(\theta)$ will be close to its population version

$$M(\theta) := P[m_\theta(X)], \qquad \theta \in \Theta.$$

Let

$$\theta_0 := \arg\max_{\theta\in\Theta} M(\theta).$$

If $\mathbb{M}_n$ and M are uniformly close, then maybe their argmax's $\hat{\theta}_n$ and θ_0 are also close. A key tool to studying such behavior of $\hat{\theta}_n$ is the *argmax continuous mapping theorem* which we consider next. Before we present the result in a general setup let us discuss the main idea behind the proof. For any given $\epsilon > 0$, we have to bound the probability $\mathbb{P}(d(\hat{\theta}_n, \theta_0) \geq \epsilon)$. The key step is to realize that

$$\begin{aligned} \mathbb{P}(d(\hat{\theta}_n,\theta_0) \geq \epsilon) \;&\leq\; \mathbb{P}\left(\sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\epsilon} [\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)] > 0\right) \\ \leq\; \mathbb{P}&\left(\sup_{\theta\in\Theta: d(\theta,\theta_0)\geq\epsilon} [(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta_0)] > -\sup_{d(\theta,\theta_0)\geq\epsilon}[M(\theta) - M(\theta_0)]\right). \end{aligned} \tag{163}$$

The (uniform) closeness of $\mathbb{M}_n$ and M (cf. condition (3) in Theorem 12.1 below) shows that the left-hand side of (163) must converge to 0 (in probability), whereas if M has a *well-separated* unique maximum[121] (cf. condition (1) in Theorem 12.1) then the right-hand side of (163) must exceed a positive number, thereby showing that $\mathbb{P}(d(\hat{\theta}_n,\theta_0) \geq \epsilon) \to 0$ as $n \to \infty$. This was carried out in Subsection 3.5.1 while discussing the consistency of M-estimators.

[121]i.e., the function $M(\theta)$ should be strictly smaller than $M(\theta_0)$ on the complement of every neighborhood of the point θ_0.

12.1 Argmax continuous mapping theorems

We state our first argmax continuous mapping theorem below which generalizes the above discussed setup (so that it can also be used to derive asymptotic distributions of the M-estimator). Our first result essentially says that the argmax functional is *continuous* at functions M that have a well-separated unique maximum.

Theorem 12.1. *Let H be a metric space and let $\{\mathbb{M}_n(h), h \in H\}$ and $\{M(h), h \in H\}$ be stochastic processes indexed by H. Suppose the following conditions hold:*

1. *$\hat{h}$ is a random element of H which satisfies*

$$M(\hat{h}) > \sup_{h \notin G} M(h) \qquad a.s.,$$

for every open set G containing $\hat{h}$; i.e., M has a unique "well-separated" point of maximum.

2. *For each n, let $\hat{h}_n \in H$ satisfy*

$$\mathbb{M}_n(\hat{h}_n) \geq \sup_{h \in H} \mathbb{M}_n(h) - o_{\mathbb{P}}(1).$$

3. *$\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(H)$.*

Then $\hat{h}_n \xrightarrow{d} \hat{h}$ in H.

Proof. By the Portmanteau theorem 10.5, to prove $\hat{h}_n \xrightarrow{d} \hat{h}$ it suffices to show that

$$\limsup_{n \to \infty} \mathbb{P}^*\{\hat{h}_n \in F\} \leq \mathbb{P}\{\hat{h} \in F\} \tag{164}$$

for every closed subset F of H. Fix a closed set F and note that

$$\{\hat{h}_n \in F\} \subseteq \left\{ \sup_{h \in F} \mathbb{M}_n(h) \geq \sup_{h \in H} \mathbb{M}_n(h) - o_{\mathbb{P}}(1) \right\}.$$

Therefore,

$$\mathbb{P}^*\big(\hat{h}_n \in F\big) \leq \mathbb{P}^*\Big(\sup_{h \in F} \mathbb{M}_n(h) - \sup_{h \in H} \mathbb{M}_n(h) + o_{\mathbb{P}}(1) \geq 0 \Big).$$

The map $\sup_{h \in F} \mathbb{M}_n(h) - \sup_{h \in H} \mathbb{M}_n(h)$ converges in distribution to $\sup_{h \in F} M(h) - \sup_{h \in H} M(h)$ as $\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(H)$ and by the continuous mapping theorem. We thus have

$$\limsup_{n \to \infty} \mathbb{P}^*\Big(\hat{h}_n \in F\Big) \leq \mathbb{P}\left(\sup_{h \in F} M(h) \geq \sup_{h \in H} M(h) \right),$$

where we have again used the Portmanteau theorem. The first assumption of the theorem implies that $\{\sup_{h \in F} M(h) \geq \sup_{h \in H} M(h)\} \subseteq \{\hat{h} \in F\}$ (note that F^c is open). This proves (164). $\square$

The idea behind the proof of the above theorem can be used to prove the following stronger technical lemma.

Lemma 12.2. *Let H be a metric space and let $\{\mathbb{M}_n(h) : h \in H\}$ and $\{M(h) : h \in H\}$ be stochastic processes indexed by H. Let A and B be arbitrary subsets of H. Suppose the following conditions hold:*

1. *$\hat{h}$ is a random element of H which satisfies $M(\hat{h}) > \sup_{h \in A \cap G^c} M(h)$ almost surely for every open set G containing $\hat{h}$.*
2. *For each n, let $\hat{h}_n \in H$ be such that $\mathbb{M}_n(\hat{h}_n) \geq \sup_{h \in H} \mathbb{M}_n(h) - o_{\mathbb{P}}(1)$.*
3. *$\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(A \cup B)$.*

Then

$$\limsup_{n \to \infty} \mathbb{P}^* \left(\hat{h}_n \in F \cap A \right) \leq \mathbb{P}\left(\hat{h} \in F \right) + \mathbb{P}\left(\hat{h} \in B^c \right) \tag{165}$$

for every closed set F.

Observe that Theorem 12.1 is a special case of this lemma which corresponds to $A = B = H$.

Proof of Lemma 12.2. The proof is very similar to that of Theorem 12.1. Observe first that

$$\left\{ \hat{h}_n \in F \cap A \right\} \subseteq \left\{ \sup_{h \in F \cap A} \mathbb{M}_n(h) - \sup_{h \in B} \mathbb{M}_n(h) + o_P(1) \geq 0 \right\}.$$

The term $\sup_{h \in F \cap A} \mathbb{M}_n(h) - \sup_{h \in B} \mathbb{M}_n(h) + o_P(1)$ converges in distribution to $\sup_{h \in F \cap A} M(h) - \sup_{h \in B} M(h)$ because $\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(A \cup B)$. This therefore gives

$$\limsup_{n \to \infty} \mathbb{P}^* \left(\hat{h}_n \in F \cap A \right) \leq \mathbb{P} \left(\sup_{h \in F \cap A} M(h) - \sup_{h \in B} M(h) \geq 0 \right)$$

Now if the event $\{\sup_{h \in F \cap A} M(h) \geq \sup_{h \in B} M(h)\}$ holds and if $\hat{h} \in B$, then $\sup_{h \in F \cap A} M(h) \geq M(\hat{h})$ which can only happen if $\hat{h} \in F$. This means

$$\mathbb{P} \left(\sup_{h \in F \cap A} M(h) - \sup_{h \in B} M(h) \geq 0 \right) \leq \mathbb{P}(\hat{h} \in B^c) + \mathbb{P}(\hat{h} \in F)$$

which completes the proof. $\square$

We next prove a more applicable argmax continuous mapping theorem. The assumption that $\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(H)$ is too stringent. It is much more reasonable to assume that $\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(K)$ for every *compact* subset K of H. The next theorem proves that $\hat{h}_n$ converges in law to $\hat{h}$ under this weaker assumption.

As we will be restricting analysis to compact sets in the next theorem, we need to assume that $\hat{h}_n$ and $\hat{h}$ lie in compact sets with arbitrarily large probability. This condition, made precise below, will be referred to as the *tightness condition*:

For every $\epsilon > 0$, there exists a compact set $K_\epsilon \subseteq H$ such that

$$\limsup_{n\to\infty} \mathbb{P}^*\Big(\hat{h}_n \notin K_\epsilon\Big) \le \epsilon \quad \text{and} \quad \mathbb{P}\Big(\hat{h} \notin K_\epsilon\Big) \le \epsilon. \tag{166}$$

Theorem 12.3 (Argmax continuous mapping theorem)**.** *Let H be a metric space and let $\{\mathbb{M}_n(h) : h \in H\}$ and $\{M(h) : h \in H\}$ be stochastic processes indexed by H. Suppose that the following conditions hold:*

1. *$\mathbb{M}_n \xrightarrow{d} M$ in $\ell^\infty(K)$ for every compact subset K of H.*
2. *Almost all sample paths $h \mapsto M(h)$ are upper semicontinuous*[122] *(u.s.c.) and possess a unique maximum at a random point $\hat{h}$.*
3. *For each n, let $\hat{h}_n$ be a random element of H such that $\mathbb{M}_n(\hat{h}_n) \ge \sup_{h\in H} \mathbb{M}_n(h) - o_{\mathbb{P}}(1)$.*
4. *The tightness condition (166) holds.*

Then $\hat{h}_n \xrightarrow{d} \hat{h}$ in H.

Proof. Let K be an arbitrary compact subset of H. We first claim that

$$M(\hat{h}) > \sup_{h\in K\cap G^c} M(h)$$

for every open set G containing $\hat{h}$. Suppose, for the sake of contradiction, that $M(\hat{h}) = \sup_{h\in K\cap G^c} M(h)$ for some open set G containing $\hat{h}$. In that case, there exist $h_m \in K \cap G^c$ with $M(h_m) \to M(h)$ as $m \to \infty$. Because $K \cap G^c$ (intersection of a closed set with a compact set) is compact, a subsequence of $\{h_m\}$ converges which means that we can assume, without loss of generality, that $h_m \to h$ for some $h \in K \cap G^c$. By the u.s.c. hypothesis, this implies that $\limsup_{m\to\infty} M(h_m) \le M(h)$ which is same as $M(\hat{h}) \le M(h)$. This implies that $\hat{h}$ is not a unique maximum (as $\hat{h} \in G$ and $h \in G^c$, we note that $\hat{h} \neq h$). This proves the claim.

We now use Lemma 12.2 with $A = B = K$ (note that $\mathbb{M}_n \xrightarrow{d} M$ on $\ell^\infty(A\cup B) = \ell^\infty(K)$). This gives that for every closed set F, we have

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}^*\Big(\hat{h}_n \in F\Big) &\le \limsup_{n\to\infty} \mathbb{P}^*\Big(\hat{h}_n \in F \cap K\Big) + \limsup_{n\to\infty} \mathbb{P}^*\Big(\hat{h}_n \in K^c\Big) \\ &\le \mathbb{P}\Big(\hat{h} \in F\Big) + \mathbb{P}\Big(\hat{h} \in K^c\Big) + \limsup_{n\to\infty} \mathbb{P}^*\Big(\hat{h}_n \in K^c\Big).\end{aligned}$$

The term on the right hand side above can be made smaller than $\mathbb{P}(\hat{h} \in F) + \epsilon$ for every $\epsilon > 0$ by choosing K appropriately (using tightness). An application of the Portmanteau theorem now completes the proof. ∎

[122]Recall the definition of upper semicontinuity: f is u.s.c. at x_0 if $\limsup_{n\to\infty} f(x_n) \le f(x_0)$ whenever $x_n \to x_0$ as $n \to \infty$.

As a simple consequence of Theorems 12.1 and 12.3, we can prove the following theorem which is useful for checking consistency of M-estimators. Note that $\mathbb{M}_n \xrightarrow{d} M$ for a deterministic process M is equivalent to $\mathbb{M}_n \xrightarrow{\mathbb{P}} M$. This latter statement is equivalent to $\sup_{h\in H} |\mathbb{M}_n(h) - M(h)|$ converges to 0 in probability.

Theorem 12.4 (Consistency Theorem)**.** *Let Θ be a metric space. For each $n \geq 1$, let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a stochastic process. Also let $\{M(\theta) : \theta \in \Theta\}$ be a deterministic process.*

1. *Suppose $\sup_{\theta\in\Theta} |\mathbb{M}_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Also suppose the existence of $\theta_0 \in \Theta$ such that $M(\theta_0) > \sup_{\theta\notin G} M(\theta)$ for every open set G containing θ_0. Then any sequence sequence of M-estimators $\hat{\theta}_n$ (assuming that $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta} \mathbb{M}_n(\theta) - o_P(1)$ is enough), converges in probability to θ_0.*

2. *Suppose $\sup_{\theta\in K} |\mathbb{M}_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ for every compact subset K of Θ. Suppose also that the deterministic limit process M is upper semicontinuous and has a unique maximum at θ_0. If $\{\hat{\theta}_n\}$ is tight, then $\hat{\theta}_n$ converges to θ_0 in probability.*

Remark 12.1. *For M-estimators, we can apply the above theorem with $\mathbb{M}_n(\theta) := \sum_{i=1}^n m_\theta(X_i)/n$ and $M(\theta) := P[m_\theta]$. In this case, the condition $\sup_{\theta\in K} |\mathbb{M}_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$ is equivalent to $\{m_\theta : \theta \in K\}$ being P-Glivenko-Cantelli.*

Theorem 12.3 can also be used to prove asymptotic distribution results for M-estimators, as illustrated in the following examples.

12.2 Asymptotic distribution

In this section we present one result that gives the asymptotic distribution of M-estimators for the case of i.i.d. observations. The formulation is from [van der Vaart, 1998]. The limit distribution of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ follows from the following theorem, where $\hat{\theta}_n$ is an M-estimator of the finite dimensional parameter θ_0 (i.e., $\hat{\theta}_n := \arg\max_{\theta\in\Theta} \mathbb{M}_n(\theta)$ where $\mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta(X)]$).

Example 12.5 (Parametric maximum likelihood estimators)**.** *Suppose $X_1, \ldots, X_n$ are i.i.d. from an unknown density p_{θ_0} belonging to a known class $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. Let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ_0. A classical result is that, under some smoothness assumptions, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N_k(0, I^{-1}(\theta_0))$ where $I(\theta_0)$ denotes the Fisher information matrix.*

This result can be derived from the argmax continuous mapping theorem. The first step is to observe that if $\theta \mapsto p_\theta(x)$ is sufficiently smooth at θ_0, then, for any $h \in \mathbb{R}^k$,

$$\sum_{i=1}^n \log \frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} = h^\top \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^\top I(\theta_0) h + o_{P_{\theta_0}}(1) \tag{167}$$

where $\dot{\ell}_{\theta_0}(x) := \nabla_\theta \log p_\theta(x)$ denotes the score function. Condition (167) is known as the LAN (local asymptotic normality) condition. We shall prove the asymptotic normality of $\hat{\theta}_n$ assuming the marginal convergence of (167) (for every fixed h) can be suitably strengthened to a process level result in $\ell^\infty(K)$, for $K \subset \mathbb{R}^k$ compact. We apply the argmax continuous mapping theorem (Theorem 12.3) with $H = \mathbb{R}^k$,

$$\mathbb{M}_n(h) := \sum_{i=1}^{n} \log \frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \quad \text{and} \quad M(h) := h^T\Delta - \frac{1}{2}h^T I(\theta_0)h$$

where $\Delta \sim N_k(0, I(\theta_0))$. Then $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\hat{h} \sim N(0, I^{-1}(\theta_0))$. The argmax theorem will then imply the result provided the conditions of the argmax theorem hold. The main condition is tightness of $\{\hat{h}_n\}$ which means that the rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{-1/2}$.

The above idea can be easily extended to derive the asymptotic distributions of other $\sqrt{n}$-consistent estimators, e.g., non-linear regression, robust regression, etc. (see [van der Vaart, 1998, Chapter 5] for more details).

Theorem 12.6. *Suppose that $x \mapsto m_\theta(x)$ is a measurable function for each $\theta \in \Theta \subset \mathbb{R}^d$ for an open set Θ, that $\theta \mapsto m_\theta(x)$ is differentiable at $\theta_0 \in \Theta$ for P-almost every x with derivative $\dot{m}_{\theta_0}(x)$, and that*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq F(x)\|\theta_1 - \theta_2\| \tag{168}$$

holds for all θ_1, θ_2 in a neighborhood of θ_0, where $F \in L_2(P)$. Also suppose that $M(\theta) = P[m_\theta]$ has a second order Taylor expansion

$$P[m_\theta] - P[m_{\theta_0}] = \frac{1}{2}(\theta - \theta_0)^\top V(\theta - \theta_0) + o(\|\theta - \theta_0\|^2)$$

where θ_0 is a point of maximum of M and V is symmetric and nonsingular (negative definite since M is a maximum at θ_0). If $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_{\mathbb{P}}(n^{-1})$ and $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) + o_{\mathbb{P}}(1) \xrightarrow{d} N_d(0, V^{-1}P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top]V^{-1}),$$

where $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P)$.

Proof. We will show that

$$\tilde{\mathbb{M}}_n(h) := n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) \xrightarrow{d} h^\top\mathbb{G}(\dot{m}_{\theta_0}) + \frac{1}{2}h^\top Vh =: \mathbb{M}(h) \text{ in } \ell^\infty(\{h : \|h\| \leq K\})$$

for every $K > 0$. Then the conclusion follows from the argmax continuous Theorem 12.3 upon noticing that

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, \mathbb{M}(h) = -B^{-1}\mathbb{G}(\dot{m}_{\theta_0}) \sim N_d(0, V^{-1}P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top)V^{-1}).$$

Now, observe that

$$n\mathbb{P}_n(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) = \sqrt{n}(\mathbb{P}_n - P)[\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0})] + nP(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}).$$

By the second order Taylor expansion of $M(\theta) := P[m_\theta]$ about θ_0, the second term of the right side of the last display converges to $(1/2)h^\top Vh$ uniformly for $\|h\| \le K$. To handle the first term we use the Donsker theorem with chaining classes. The classes

$$\mathcal{F}_n := \{\sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}) : \|h\| \le K\}$$

have envelopes $F_n = F = \dot{m}_{\theta_0}$ for all n, and since $\dot{m}_{\theta_0} \in L_2(P)$ the Lindeberg condition is satisfied easily. Furthermore, with

$$f_{n,g} = \sqrt{n}(m_{\theta_0+gn^{-1/2}} - m_{\theta_0}), \qquad f_{n,h} = \sqrt{n}(m_{\theta_0+hn^{-1/2}} - m_{\theta_0}),$$

by the dominated convergence theorem the covariance functions satisfy

$$P(f_{n,g}f_{n,h}) - P(f_{n,g})P(f_{n,h}) \to P(g^\top \dot{m}_{\theta_0}\dot{m}_{\theta_0}^\top h) = g^\top \mathbb{E}[\mathbb{G}(\dot{m}_{\theta_0})\mathbb{G}(\dot{m}_{\theta_0}^\top)]h.$$

Finally, the bracketing entropy condition holds since, by way of the same entropy calculations used in the proof of we have

$$N_{[\,]}(2\epsilon\|F\|_{P,2}, \mathcal{F}_n, L_2(P)) \le \left(\frac{CK}{\epsilon}\right)^d, \quad \text{i.e.,}\;\; N_{[\,]}(\epsilon, \mathcal{F}_n, L_2(P)) \lesssim \left(\frac{CK\|F\|_{P,2}}{\epsilon}\right)^d$$

Thus, $J_{[\,]}(\delta, \mathcal{F}_n, L_2(P)) \lesssim \int_0^\delta \sqrt{d \log\left(\frac{CK}{\epsilon}\right)} d\epsilon$, and hence the bracketing entropy hypothesis of Donsker theorem holds. We conclude that $\tilde{\mathbb{M}}_n(h)$ converges weakly to $h^\top \mathbb{G}(\dot{m}_{\theta_0})$ in $\ell^\infty(\{h : \|h\| \le K\})$, and the desired result holds. $\square$

12.3 A non-standard example

Example 12.7 (Analysis of the shorth)**.** *Recall the setup of Example 5.4. Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with density p with respect to the Lebesgue measure. Let F_X be the distribution function of X. Suppose that p is a unimodal symmetric density with mode θ_0 (with $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$). We want to estimate θ_0.*

Let

$$\mathbb{M}(\theta) := P[m_\theta] = \mathbb{P}(|X - \theta| \le 1) = F_X(\theta + 1) - F_X(\theta - 1)$$

where $m_\theta(x) = \mathbf{1}_{[\theta-1,\theta+1]}(x)$. We can how that $\theta_0 = \operatorname{argmax}_{\theta\in\mathbb{R}} \mathbb{M}(\theta)$.

We can estimate θ_0 by

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta\in\mathbb{R}} \mathbb{M}_n(\theta), \qquad \textit{where} \qquad \mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta.$$

We have already seen that (in Example 5.4) $\tau_n := n^{1/3}(\hat{\theta}_n - \theta_0) = O_p(1)$. Let us here give a sketch of the limiting distribution of (the normalized version of) $\hat{\theta}_n$. Observe that

$$\tau_n = \operatorname*{argmax}_{h\in\mathbb{R}} \mathbb{M}_n(\theta_0 + hn^{-1/3}) = \operatorname*{argmax}_{h\in\mathbb{R}} n^{2/3}[\mathbb{M}_n(\theta_0 + hn^{-1/3}) - \mathbb{M}_n(\theta_0)].$$

The plan is to show that the localized (and properly normalized) stochastic process $\tilde{\mathbb{M}}_n(h) := n^{2/3}[\mathbb{M}_n(\theta_0+hn^{-1/3})-\mathbb{M}_n(\theta_0)]$ *converges in distribution to "something" so that we can apply the argmax continuous mapping theorem (Theorem 12.3) to deduce the limiting behavior of* τ_n. *Notice that,*

$$\begin{aligned}\tilde{\mathbb{M}}_n(h) &:= n^{2/3}\mathbb{P}_n[m_{\theta_0+hn^{-1/3}} - m_{\theta_0}] \\ &= n^{2/3}(\mathbb{P}_n - P)[m_{\theta_0+hn^{-1/3}} - m_{\theta_0}] + n^{2/3}P[m_{\theta_0+hn^{-1/3}} - m_{\theta_0}],\end{aligned}$$

where the second term is

$$\begin{aligned}n^{2/3}[\mathbb{M}(\theta_0+hn^{-1/3}) - \mathbb{M}(\theta_0)] &= n^{2/3}\mathbb{M}'(\theta_0)hn^{-1/3} + n^{2/3}\frac{1}{2}\mathbb{M}''(\theta^*)h^2n^{-2/3} \\ &\to \frac{1}{2}\mathbb{M}''(\theta_0)h^2 = \frac{1}{2}[p'(\theta_0+1) - p'(\theta_0-1)]h^2,\end{aligned}$$

uniformly in $|h| \le K$, *for any constant* K. *Note that as* $\mathbb{M}$ *is differentiable* $\mathbb{M}'(\theta_0) = p(\theta_0+1) - p(\theta_0-1) = 0$. *Thus, we want to study the empirical process* $\mathbb{G}_n$ *indexed by the collection of functions* $\mathcal{F}_n := \{n^{1/6}(m_{\theta_0+hn^{-1/3}} - m_{\theta_0}) : |h| \le K\}$. *Here we can apply a Donsker theorem for a family of functions depending on n, for example Theorem 11.7. Thus we need to check that* (160) *and* (161) *hold. Observe that*

$$\begin{aligned}&P[(f_{n,s} - f_{n,t})^2] - [P(f_{n,s} - f_{n,t})]^2 \\ &= n^{1/3}P[(m_{\theta_0+sn^{-1/3}} - m_{\theta_0+tn^{-1/3}})^2] - o(1) \\ &= n^{1/3}\Big\{P\mathbf{1}_{[\theta_0-1+sn^{-1/6},\theta_0-1+tn^{-1/6}]} + P\mathbf{1}_{[\theta_0+1+sn^{-1/6},\theta_0+1+tn^{-1/6}]}\Big\} + o(1) \qquad \text{if} \quad t > s \\ &\to [p(\theta_0-1) + p(\theta_0+1)]|s-t|.\end{aligned}$$

Thus, we can conclude that

$$n^{2/3}(\mathbb{P}_n - P)[m_{\theta_0+hn^{-1/3}} - m_{\theta_0}] \xrightarrow{d} a\mathcal{Z}(h)$$

where $a^2 := p(\theta_0+1) + p(\theta_0-1)$ *and* $\mathcal{Z}$ *is a standard two-sided Brownian motion process starting from* 0 *(show this!). We can now use the argmax continuous mapping theorem to conclude now that*

$$\tau_n = n^{1/3}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmax}_h[a\mathcal{Z}(h) - bh^2],$$

where $b := -\mathbb{M}''(\theta_0)/2$.

13 Concentration Inequalities

We are interested in bounding the random fluctuations of (complicated) functions of many independent random variables. Let $X_1, \ldots, X_n$ be *independent* random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$, and let

$$Z = f(X_1, \ldots, X_n)$$

be the random variable of interest (e.g., $Z = \sum_{i=1}^n X_i$). In this section we try to understand: (a) under what conditions does Z concentrate around its mean $\mathbb{E}Z$? (b) how large are typical deviations of Z from $\mathbb{E}Z$? In particular, we seek upper bounds for

$$\mathbb{P}(Z > \mathbb{E}Z + t) \qquad \text{and} \qquad \mathbb{P}(Z < \mathbb{E}Z - t) \qquad \text{for } t > 0.$$

Various approaches have been used over the years to tackle such questions, including martingale methods, information theoretic methods, logarithmic Sobolev inequalities, etc. We have already seen many concentration inequalities in this course, e.g., Hoeffding's inequality, Bernstein's inequality, Talagrand's concentration inequality, etc.

Let $X_1, \ldots, X_n$ be *independent* random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and let

$$Z = f(X_1, \ldots, X_n)$$

be the random variable of interest. Recall the martingale representation of $Z = \sum_{i=1}^n \Delta_i$ and the notation introduced in Section 3.6. Note that,

$$\mathrm{Var}(Z) = \mathbb{E}\left[\Big(\sum_{i=1}^n \Delta_i\Big)^2\right] = \sum_{i=1}^n \mathbb{E}(\Delta_i^2) + 2\sum_{j>i} \mathbb{E}(\Delta_i \Delta_j).$$

Now if $j > i$, $\mathbb{E}_i \Delta_j = \mathbb{E}_i(\mathbb{E}_j Z) - \mathbb{E}_i(\mathbb{E}_{j-1} Z) = \mathbb{E}_i(Z) - \mathbb{E}_i(Z) = 0$, so

$$\mathbb{E}_i(\Delta_j \Delta_i) = \Delta_i \mathbb{E}_i(\Delta_j) = 0.$$

Thus, we obtain that

$$\mathrm{Var}(Z) = \mathbb{E}\left[\Big(\sum_{i=1}^n \Delta_i\Big)^2\right] = \sum_{i=1}^n \mathbb{E}(\Delta_i^2). \tag{169}$$

13.1 Efron-Stein inequality

Until now we have not made any use of the fact that Z is a function of independent variables. Indeed,

$$\mathbb{E}_i Z = \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_i, x_{i+1}, \ldots, x_n) d\mu_{i+1}(x_{i+1}) \ldots d\mu_n(x_n),$$

where, for every $j = 1, \ldots, n$, μ_j denotes the probability distribution of X_j.

Let $\mathbb{E}^{(i)}(Z)$ denote the expectation of Z with respect to the i-th variable X_i only, fixing the values of the other variables, i.e.,

$$\mathbb{E}^{(i)}(Z) = \int_{\mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n) d\mu_i(x_i).$$

Using Fubini's theorem,

$$\mathbb{E}_i(\mathbb{E}^{(i)} Z) = \mathbb{E}_{i-1} Z. \tag{170}$$

Theorem 13.1 (Efron-Stein inequality)**.** *Then,*

$$\mathrm{Var}(Z) \leq \mathbb{E} \sum_{i=1}^{n} (Z - \mathbb{E}^{(i)}(Z))^2 = \mathbb{E} \sum_{i=1}^{n} \mathrm{Var}^{(i)}(Z) =: v. \tag{171}$$

Moreover, if $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$, and if we define, for every $i = 1, \ldots, n$,

$$Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n), \tag{172}$$

then

$$v = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')^2].$$

Also, for every i, letting

$$Z_i := f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

for any arbitrary measurable function f_i, we have

$$v = \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i)^2].$$

Proof. Using (170) we may write $\Delta_i = \mathbb{E}_i(Z - \mathbb{E}^{(i)} Z)$. By Jensen's inequality, used conditionally, $\Delta_i^2 \leq \mathbb{E}_i[(Z - \mathbb{E}^{(i)} Z)^2]$. Now (169) yields the desired bounds.

To see the second claim, we use (conditionally) the fact that if X and X' are i.i.d. real valued random variables, then $\mathrm{Var}(X) = \mathbb{E}[(X - X')^2]/2$. Since conditionally on $X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, Z_i' is an independent copy of Z, we may write

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2} \mathbb{E}^{(i)}[(Z - Z_i')^2],$$

where we have used the fact that the conditional distributions of Z and Z_i' are identical. The last identity is obtained by recalling that, for any real-values random variable X, $\mathrm{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$. Using this fact conditionally, we have, for every $i = 1, \ldots, n$,

$$\mathrm{Var}^{(i)}(Z) = \inf_{Z_i} \mathbb{E}^{(i)}[(Z - Z_i)^2].$$

□

Observe that in the case when $Z = \sum_{i=1}^n X_i$, the Efron-Stein inequality becomes an equality. Thus, the bound in the Efron-Stein inequality, is, in a sense, not improvable.

It is easy to see that if f has the bounded difference property (see (24)) with constants $c_1, \ldots, c_n$, then

$$\mathrm{Var}(Z) \le \frac{1}{2} \sum_{i=1}^n c_i^2.$$

Example 13.2 (Kernel density estimation)**.** *Recall Example 3.25. Let $X_1, \ldots, X_n$ are i.i.d. from a distribution P on $\mathbb{R}$ (the argument can be easily generalized to $\mathbb{R}^d$) with density ϕ. The kernel density estimator (KDE) of ϕ is $\hat{\phi}_n : \mathbb{R} \to [0, \infty)$ defined as $\hat{\phi}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)$, where $h_n > 0$ is the smoothing bandwidth and K is a non-negative kernel (i.e., $K \ge 0$ and $\int K(x)dx = 1$). The L_1-error of the estimator $\hat{\phi}_n$ is $Z := f(X_1, \ldots, X_n) := \int |\hat{\phi}_n(x) - \phi(x)| dx$. We have shown in (26) that f satisfies (24) with $c_i = \frac{2}{n}$. Thus the difference between $Z - Z_i'$, deterministically, is upper bounded by $2/n$, for all i. Thus, an application of the Efron-Stein inequality gives*

$$\mathrm{Var}(Z) \le \frac{n}{2} \left(\frac{2}{n}\right)^2 = \frac{2}{n}.$$

It is known that for every ϕ, $\sqrt{n}\mathbb{E}(Z_n) \to \infty$ (we write Z_n instead of Z to emphasize the dependence on n), which implies, by Chebyshev's inequality, for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{Z_n}{\mathbb{E}(Z_n)} - 1\right| \ge \epsilon\right) = \mathbb{P}\Big(|Z_n - \mathbb{E}(Z_n)| \ge \epsilon\, \mathbb{E}(Z_n)\Big) \le \frac{\mathrm{Var}(Z_n)}{\epsilon^2 [\mathbb{E}(Z_n)]^2} \to 0$$

as $n \to \infty$. Thus, $Z_n / \mathbb{E}(Z_n) \xrightarrow{p} 1$, or in other words, Z_n is relatively stable. This means that the random L_1-error essentially behaves like its expected value.

Example 13.3. *Let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$, and let $X_1, \ldots, X_n$ be n i.i.d random points in $\mathcal{X}$ with distribution P. Let $Z = \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - P(A)|$. By the Efron-Stein inequality (**show this**),*

$$\mathrm{Var}(Z) \le \frac{2}{n},$$

regardless of the richness of the collection of sets $\mathcal{A}$ and the distribution P.

Recall that we can bound $\mathbb{E}(Z)$ using empirical process techniques. Let $\mathbb{P}_n'(A) = \sum_{i=1}^n \mathbf{1}_A(X_i')/n$ where $X_1', \ldots, X_n'$ (sometimes called a "ghost" sample) are independent copies of $X_1, \ldots, X_n$. Let $\mathbb{E}'$ denote the expectation only with respect to $X_1', \ldots, X_n'$.

$$\begin{aligned}
& \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - P(A)| = \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{E}'[\mathbb{P}_n(A) - \mathbb{P}_n'(A)]| \\
\le \quad & \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}_n'(A)| = \frac{1}{n} \mathbb{E} \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n (\mathbf{1}_A(X_i) - \sum_{i=1}^n \mathbf{1}_A(X_i')) \right|
\end{aligned}$$

where we have used the Jensen's inequality. Next we use symmetrization: if $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables, then the last term in the previous display can be bounded (from above) by (**Exercise**)

$$\frac{1}{n}\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}(\mathbf{1}_A(X_i) - \sum_{i=1}^{n}\mathbf{1}_A(X_i'))\right| \leq \frac{2}{n}\mathbb{E}\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\epsilon_i\mathbf{1}_A(X_i)\right|.$$

Note that $\sum_{i=1}^{n}\epsilon_i\mathbf{1}_A(X_i)$ can be thought of as the sample covariance between the ϵ_i's and the $\mathbf{1}_A(X_i)$'s. Letting,

$$R_n = \frac{1}{n}\mathbb{E}_\epsilon\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\epsilon_i\mathbf{1}_A(X_i)\right|,$$

where $\mathbb{E}_\epsilon$ denotes the expectation with respect to the Rademacher variables, we see that

$$\mathbb{E}\sup_{A\in\mathcal{A}}|\mathbb{P}_n(A) - \mathbb{P}(A)| \leq 2\mathbb{E}R_n.$$

Observe that R_n is a data dependent quantity, and does not involve the probability measure P. We want to show that R_n is concentrated around its mean $\mathbb{E}R_n$. Define

$$R_n^{(i)} = \frac{1}{n}\mathbb{E}_\epsilon\sup_{A\in\mathcal{A}}\left|\sum_{j\neq i}\epsilon_j\mathbf{1}_A(X_j)\right|,$$

one can show that (**Exercise**)

$$0 \leq n(R_n - R_n^{(i)}) \leq 1, \qquad \text{and} \qquad \sum_{i=1}^{n}n(R_n - R_n^{(i)}) \leq nR_n. \tag{173}$$

By the Efron-Stein inequality,

$$\mathrm{Var}(nR_n) \leq \mathbb{E}\sum_{i=1}^{n}[n(R_n - R_n^{(i)})]^2 \leq \mathbb{E}(nR_n).$$

Random variables with the property (173) are called *self-bounding*. These random variables have their variances bounded by their means an thus are automatically concentrated.

Recall that $\Delta_n(\mathcal{A}, \mathbf{X}_n)$ is the number of different sets of the form $\{X_1, \ldots, X_n\} \cap A$ where $A \in \mathcal{A}$, then R_n is the maximum of $\Delta_n(\mathcal{A}, \mathbf{X}_n)$ sub-Gaussian random variables. By the maximal inequality,

$$R_n \leq \sqrt{\frac{\log \Delta_n(\mathcal{A}, \mathbf{X}_n)}{2n}}.$$

Let $V = V(\mathbf{x}_n, \mathcal{A})$ be the size of the largest subset of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{A}$. Note that V is a random variable. In fact, V is self-bounding. Let $V^{(i)}$ be defined similarly with all the points excluding the i'th point. Then, (**Exercise**) for every $1 \leq i \leq n$,

$$0 \leq V - V^{(i)} \leq 1 \qquad \text{and} \qquad \sum_{i=1}^{n}(V - V^{(i)}) \leq V.$$

Thus, $\sum_{i=1}^{n}(V - V^{(i)})^2 \leq V$, and so by the Efron-Stein inequality, $\mathrm{Var}(V) \leq \mathbb{E}V$.

13.2 Concentration and logarithmic Sobolev inequalities

We start with a brief summary of some basic properties of the (Shannon) entropy of a random variable. For simplicity, we will only consider discrete-values random variables.

Definition 13.4 (Shannon entropy)**.** *Let X be a random variable taking values in a countable set $\mathcal{X}$ with probability mass function (p.m.f.) $p(x) = \mathbb{P}(X = x), x \in \mathcal{X}$. The Shannon entropy (or just entropy) of X is defined as*

$$H(X) := \mathbb{E}[-\log p(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{174}$$

where log *denotes natural logarithm and $0 \log 0 = 0$. The entropy can be thought of as the "uncertainty" in the random variable. Observe that the entropy is obviously nonnegative.*

If X and Y is a pair of discrete random variables taking values in $\mathcal{X} \times \mathcal{Y}$ then the *joint entropy* $H(X, Y)$ of X and Y is defined as the entropy of the pair (X, Y).

Definition 13.5 (Kullback-Leibler divergence)**.** *Let P and Q be two probability distributions with p.m.f.'s p and q. Then the Kullback-Leibler divergence or relative entropy of P and Q is*

$$D(P\|Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

if P is absolutely continuous with respect to Q and infinite otherwise.

It can be shown that $D(P\|Q) \geq 0$, and $D(P\|Q) = 0$ if and only if $P = Q$. This follows from observing the fact that if P is absolutely continuous with respect to Q, since $\log x \leq x - 1$ for all $x > 0$,

$$D(P\|Q) = -\sum_{x \in \mathcal{X}: p(x) > 0} p(x) \log \frac{q(x)}{p(x)} \geq \sum_{x \in \mathcal{X}: p(x) > 0} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \geq 0.$$

Observe that if $X \sim P$ takes N values, then by taking Q to be the uniform distribution on those N values,

$$D(P\|Q) = \log N - H(X), \quad \text{and} \quad H(X) \leq \log N \quad (\text{as } D(P\|Q) \geq 0).$$

Definition 13.6 (Conditional entropy)**.** *Consider a pair of random variables (X, Y). The conditional entropy $H(X|Y)$ is defined as*

$$H(X|Y) = H(X, Y) - H(Y).$$

Observe that if we write $p(x, y) = \mathbb{P}(X = x, Y = y)$ and $p(x|y) = \mathbb{P}(X = x|Y = y)$, then

$$H(X|Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x|y) = \mathbb{E}[-\log p(X|Y)].$$

It is easy to see that $H(X|Y) \geq 0$.

Suppose that $(X, Y) \sim P_{X,Y}$ and $X \sim P_X$ and $Y \sim P_Y$. Noting that $D(P_{X,Y} \| P_X \otimes P_Y) = H(X) - H(X|Y)$, the nonnegativity of the relative entropy implies

$$H(X|Y) \leq H(X). \tag{175}$$

The *chain rule* for entropy says that for random variables $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1}). \tag{176}$$

Let $X = (X_1, \ldots, X_n)$ be a vector of n random variables (not necessarily independent) and let $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ (vector obtained by leaving out X_i). We have the following result.

Theorem 13.7 (Han's inequality)**.** *Let $X_1, \ldots, X_n$ be discrete random variables. Then,*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n). \tag{177}$$

Proof. Observe that

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\leq H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

where we have used (175) conditionally. Now, by summing over all i and using (176), we get the desired result. □

Next we derive an inequality (which may be regarded as a version of Han's inequality for relative entropies) that is fundamental in deriving a "sub-additivity" inequality (see Theorem 13.11), which, in turn, is at the basis of many exponential concentration inequalities.

Let $\mathcal{X}$ be a countable set, and let P and Q be probability measures on $\mathcal{X}^n$ such that $P = P_1 \otimes \ldots \otimes P_n$ is a product measure. We denote the elements of $\mathcal{X}^n$ by $x = (x_1, \ldots, x_n)$and write $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ for the $(n-1)$-vector obtained by leaving out the i-th component of x. Denote by $Q^{(i)}$ and $P^{(i)}$ the marginal distributions of Q and P, and let $p^{(i)}$ and $q^{(i)}$ denote the corresponding p.m.f.'s, i.e.,

$$\begin{aligned} q^{(i)}(x^{(i)}) &= \sum_{y \in \mathcal{X}} q(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \quad \text{and} \\ p^{(i)}(x^{(i)}) &= p_1(x_1) \cdots p_{i-1}(x_{i-1}) p_{i+1}(x_{i+1}) \cdots p_n(x_n). \end{aligned}$$

Then we have the following result, proved in [Boucheron et al., 2013, Section 4.6].

Theorem 13.8 (Han's inequality for relative entropies)**.**

$$D(Q \| P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)} \| P^{(i)})$$

or equivalently,

$$D(Q\|P) \le \sum_{i=1}^{n} \left[D(Q\|P) - D(Q^{(i)}\|P^{(i)}) \right].$$

We end this section with a result that will be useful in developing the entropy method, explained in the next section.

Theorem 13.9 (The expected value minimizes expected Bergman divergence)**.** *Let $I \subset \mathbb{R}$ be an open interval and let $g : I \to \mathbb{R}$ be convex and differentiable. For any $x, y \in I$, the Bergman divergence of f from x to y is $g(y) - g(x) - g'(x)(y - x)$. Let X be an I-valued random variable. Then,*

$$\mathbb{E}[g(X) - g(\mathbb{E}X)] = \inf_{a \in I} \mathbb{E}[g(X) - g(a) - g'(a)(X - a)]. \tag{178}$$

Proof. Let $a \in I$. The expected Bergman divergence from a is $\mathbb{E}[g(X)-g(a)-g'(a)(X-a)]$. The expected Bergman divergence from $\mathbb{E}X$ is

$$\mathbb{E}[g(X) - g(\mathbb{E}X) - g'(\mathbb{E}X)(X - \mathbb{E}X)] = \mathbb{E}[g(X) - g(\mathbb{E}X)].$$

Thus, the difference between the expected Bergman divergence from a and the expected from $\mathbb{E}X$ is

$$\begin{aligned} & \mathbb{E}[g(X) - g(a) - g'(a)(X - a)] - \mathbb{E}[g(X) - g(\mathbb{E}X)] \\ = \quad & \mathbb{E}[-g(a) - g'(a)(X - a) + g(\mathbb{E}X)] \\ = \quad & g(\mathbb{E}X) - g(a) - g'(a)(\mathbb{E}X - a) \ge 0 \end{aligned}$$

as g is convex.

□

13.3 The Entropy method

Definition 13.10 (Entropy)**.** *The entropy of a random variable $Y \ge 0$ is*

$$\mathrm{Ent}(Y) = \mathbb{E}\Phi(Y) - \Phi(\mathbb{E}Y) \tag{179}$$

where $\Phi(x) = x \log x$ for $x > 0$ and $\Phi(0) = 0$. By Jensen's inequality, $\mathrm{Ent}(Y) \ge 0$.

Remark 13.1. *Taking $g(x) = x \log x$ in (178) we obtain the following variational formula for entropy:*

$$\mathrm{Ent}(Y) = \inf_{u > 0} \mathbb{E}\left[Y(\log Y - \log u) - (Y - u) \right]. \tag{180}$$

Let $X_1, \ldots, X_n$ be independent and let $Z = f(X_1, \ldots, X_n)$, where $f \ge 0$. Denote

$$\mathrm{Ent}^{(i)}(Z) := \mathbb{E}^{(i)}[\Phi(Z)] - \Phi(\mathbb{E}^{(i)} Z).$$

Theorem 13.11 (Sub-additivity of the Entropy)**.** *Let $X_1, \dots, X_n$ be independent and let $Z = f(X_1, \dots, X_n)$, where $f \geq 0$. Then*

$$\operatorname{Ent}(Z) \leq \mathbb{E} \sum_{i=1}^{n} \operatorname{Ent}^{(i)}(Z). \tag{181}$$

Proof. (Han's inequality implies the sub-additivity property.) First observe that if the inequality is true for a random variable Z, then it is also true for cZ, where $c > 0$. Hence we may assume that $\mathbb{E}(Z) = 1$. Now define the probability measure Q on $\mathcal{X}^n$ by its p.m.f. q given by

$$q(x) = f(x)p(x), \qquad \text{for all } x \in \mathcal{X}^n$$

where p denotes the p.m.f. of $X = (X_1, \dots, X_n)$. Let P be the distribution induced by p. Then,

$$\mathbb{E}\Phi(Z) - \Phi(\mathbb{E}Z) = \mathbb{E}[Z \log Z] = D(Q\|P)$$

which, by Theorem 13.8, does not exceed $\sum_{i=1}^{n}[D(Q\|P) - D(Q^{(i)}\|P^{(i)})]$. However, straightforward calculations show that

$$\sum_{i=1}^{n}[D(Q\|P) - D(Q^{(i)}\|P^{(i)})] = \mathbb{E} \sum_{i=1}^{n} \operatorname{Ent}^{(i)}(Z)$$

and the statement follows. $\square$

Remark 13.2. *The form of the above inequality should remind us of the Efron-Stein inequality. In fact, if we take $\Phi(x) = x^2$, then the above display is the exact analogue of the Efron-Stein inequality.*

Remark 13.3 (A logarithmic Sobolev inequality on the hypercube)**.** *Let $X = (X_1, \dots, X_n)$ be uniformly distributed over $\{-1, +1\}^n$. If $f : \{-1, +1\}^n \to \mathbb{R}$ and $Z = f(X)$, then*

$$\operatorname{Ent}(Z^2) \leq \frac{1}{2}\mathbb{E} \sum_{i=1}^{n} (Z - Z_i')^2$$

where Z_i' is defined as in (172).

The proof uses the sub-additivity of the entropy and calculus for the case $n = 1$; see e.g., [Boucheron et al., 2013, Theorem 5.1]. In particular, it implies the Efron-Stein inequality.

13.4 Gaussian concentration inequality

Theorem 13.12 (Gaussian logarithmic Sobolev inequality)**.** *Let $X = (X_1, \dots, X_n)$ be a vector of n i.i.d. standard normal random variables and let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then*

$$\operatorname{Ent}(g^2) \leq 2\mathbb{E}\left[\|\nabla g(X)\|^2\right].$$

This result can be proved using the central limit theorem and the Bernoulli log-Sobelev inequality; see e.g., [Boucheron et al., 2013, Theorem 5.4].

Theorem 13.13 (Gaussian Concentration: the Tsirelson-Ibragimov-Sudakov inequality)**.** *Let $X = (X_1, \ldots, X_n)$ be a vector of n i.i.d. standard normal random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a an L-Lipschitz function, i.e., there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^n$,*

$$|f(x) - f(y)| \le L\|x - y\|.$$

Then, for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\left[e^{\lambda(f(X) - \mathbb{E}f(X))}\right] \le \frac{\lambda^2}{2}L^2,$$

and for all $t > 0$,

$$\mathbb{P}(f(X) - \mathbb{E}f(X) \ge t) \le e^{-t^2/(2L^2)}.$$

Proof. By a standard density argument we may assume that f is differentiable with gradient uniformly bounded by L. Using Theorem 13.12 for the function $e^{\lambda f/2}$, we obtain,

$$\mathrm{Ent}(e^{\lambda f}) \le 2\mathbb{E}\left[\|\nabla e^{\lambda f(X)/2}\|^2\right] = \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda f(X)}\|\nabla f(X)\|^2\right] \le \frac{\lambda^2 L^2}{2}\mathbb{E}\left[e^{\lambda f(X)}\right].$$

The Gaussian log-Sobolev inequality may now be used with

$$g(x) = e^{\lambda f(x)/2}, \qquad \text{where } \lambda \in \mathbb{R}.$$

If $F(\lambda) := \mathbb{E}e^{\lambda Z}$ is the moment generating function of $Z = f(X)$, then

$$\mathrm{Ent}(g(X)^2) = \lambda\mathbb{E}(e^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) = \lambda F'(\lambda) - F(\lambda)\log F(\lambda).$$

Thus, we obtain the differential inequality (This is usually referred to as the Herbst's argument.)

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \le \frac{\lambda^2 L^2}{2}F(\lambda). \tag{182}$$

To solve (182) divide both sides by the positive number $\lambda^2 F(\lambda)$. Defining $G(\lambda) := \log F(\lambda)$, we observe that the left-hand side is just the derivative of $G(\lambda)/\lambda$. Thus, we obtain the inequality

$$\frac{d}{d\lambda}\left(\frac{G(\lambda)}{\lambda}\right) \le \frac{L^2}{2}.$$

By l'Hospital's rule we note that $\lim_{\lambda\to 0} G(\lambda)/\lambda = F'(0)/F(0) = \mathbb{E}Z$. If $\lambda > 0$, by integrating the inequality between 0 and λ, we get $G(\lambda)/\lambda \le \mathbb{E}Z + \lambda L^2/2$, or in other words,

$$F(\lambda) \le e^{\lambda\mathbb{E}Z + \lambda^2 L^2/2}. \tag{183}$$

Finally, by Markov's inequality,

$$\mathbb{P}(Z > \mathbb{E}Z + t) \le \inf_{\lambda>0} F(\lambda)e^{-\lambda\mathbb{E}Z - \lambda t} \le \inf_{\lambda>0} e^{\lambda^2 L^2/2 - \lambda t} = e^{-t^2/(2L^2)},$$

where $\lambda = t/L^2$ minimizes the obtained upper bound. Similarly, if $\lambda < 0$, we may integrate the obtained upper bound for the derivative of $G(\lambda)/\lambda$ between $-\lambda$ and 0 to obtain the same bound as in (183), which implies the required bound for the left-tail inequality $\mathbb{P}(Z < \mathbb{E}Z - t)$. $\square$

Remark 13.4. *An important feature of the problem is that the right-hand side does not depend on the dimension n.*

Example 13.14 (Supremum of a Gaussian process)**.** *Let $(X_t)_{t\in\mathcal{T}}$ be an almost surely continuous centered Gaussian process indexed by a totally bounded set $\mathcal{T}$. Let $Z = \sup_{t\in\mathcal{T}} X_t$. If*

$$\sigma^2 := \sup_{t\in\mathcal{T}} \mathbb{E}(X_t^2),$$

then

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq u) \leq 2e^{-u^2/(2\sigma^2)}.$$

Let us first assume that $\mathcal{T}$ is a finite set (the extension to arbitrary totally bounded $\mathcal{T}$ is based on a separability argument and monotone convergence; see [Boucheron et al., 2013, Exercise 5.14]). We may assume, without loss of generality, that $\mathcal{T} = \{1, \dots, n\}$. Let Γ be the covariance matrix of the centered Gaussian vector $X = (X_1, \dots, X_n)$. Denote by A the square root of the positive semidefinite matrix Γ. If $Y = (Y_1, \dots, Y_n)$ is a vector of i.i.d. standard normal random variables, then

$$f(Y) = \max_{i=1,\dots,n} (AY)_i$$

has the same distribution as $\max_{i=1,\dots,n} X_i$. Hence we can assume the Gaussian concentration inequality by bounding the Lipchitz function f. By the Cauchy-Schwarz inequality, for all $u, v \in \mathbb{R}^n$ and $i = 1, \dots, n$,

$$|(Au)_i - (Av)_i| = \Big|\sum_j A_{ij}(u_j - v_j)\Big| \leq \Big(\sum_j A_{ij}^2\Big)^{1/2} \|u - v\|.$$

Since $\sum_j A_{ij}^2 = \mathrm{Var}(X_i)$, we get

$$|f(u) - f(v)| \leq \max_{i=1,\dots,n} |(Au)_i - (Av)_i| \leq \sigma\|u - v\|.$$

Therefore, f is Lipschitz with constant σ and the tail bound follows from the Gaussian concentration inequality.

13.5 Bounded differences inequality revisited

In the previous subsections we have used the log-Sobolev inequalities quite effectively to derive concentration results when the underlying distribution is Gaussian and Bernoulli. Here we try to extend these results to hold under more general distributional assumptions. Observe that however, (181) holds for any distribution. We state a result in that direction; see e.g., [Boucheron et al., 2013, Theorem 6.6].

Theorem 13.15 (A modified logarithmic Sobolev Inequality)**.** *Let $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$,*

$$\lambda \mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z}) \log \mathbb{E}(e^{\lambda Z}) \leq \sum_{i=1}^{n} \mathbb{E}[e^{\lambda Z}\phi(-\lambda(Z - Z_i))], \tag{184}$$

where $Z_i := f_i(X^{(i)})$ for a arbitrary function $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$.

Proof. We bound each term on the right-hand side of the sub-additivity of entropy. To do this we will apply (180) conditionally. Let Y_i be a positive function of the random variables $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, then

$$\mathbb{E}^{(i)}(Y \log Y) - (\mathbb{E}^{(i)}Y) \log(\mathbb{E}^{(i)}Y) \leq \mathbb{E}^{(i)}\left[Y(\log Y - \log Y_i) - (Y - Y_i)\right].$$

Applying the above inequality to the variables $Y = e^{\lambda Z}$ and $Y_i = e^{\lambda Z_i}$, we obtain

$$\mathbb{E}^{(i)}(Y \log Y) - (\mathbb{E}^{(i)}Y) \log(\mathbb{E}^{(i)}Y) \leq \mathbb{E}^{(i)}\left[e^{\lambda Z}\phi(-\lambda(Z - Z_i))\right].$$

□

Theorem 13.16 (A stronger form of the bounded differences inequality)**.** *Let $X_1, \ldots, X_n$ be n independent random variables, $Z = f(X_1, \ldots, X_n)$ and Z_i denotes an $X^{(i)}$-measurable random variable defined by*

$$Z_i = \inf_{x_i'} f(X_1, \ldots, X_{i-1}, x_i', X_{i+1}, \ldots, X_n). \tag{185}$$

Assume that Z is such that there exists a constant $v > 0$ for which, almost surely,

$$\sum_{i=1}^{n}(Z - Z_i)^2 \leq v.$$

Then, for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{-t^2/(2v)}.$$

Proof. The result follows from the modified logarithmic Sobolev Inequality. Observe that for $x > 0$, $\phi(-x) \leq x^2/2$. Thus, Theorem 13.15 implies

$$\lambda \mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z}) \log \mathbb{E}(e^{\lambda Z}) \leq \mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}\frac{\lambda^2}{2}(Z - Z_i)^2\right] \leq \frac{\lambda^2 v}{2}\mathbb{E}(e^{\lambda Z}).$$

The obtained inequality has the same form as (182), and the proof may be finished in an identical way (using the Herbst's argument). □

Remark 13.5. *As a consequence, if the condition $\sum_{i=1}^{n}(Z - Z_i)^2 \leq v$ is satisfied both for $Z_i = \inf_{x_i'} f(X_1, \ldots, X_{i-1}, x_i', X_{i+1}, \ldots, X_n)$ and $Z_i = \sup_{x_i'} f(X_1, \ldots, X_{i-1}, x_i', X_{i+1}, \ldots, X_n)$, one has the two-sided inequality*

$$\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq 2e^{-t^2/(2v)}.$$

Example 13.17 (The largest eigenvalue of a symmetric matrix)**.** *Let $A = (X_{ij})_{n\times n}$ be a symmetric random matrix, with X_{ij}, $i \le j$, being independent random variables (not necessarily identically distributed) with $|X_{ij}| \le 1$. Let*

$$Z = \lambda_1 := \sup_{u:\|u\|=1} u^\top A u,$$

and suppose that v is such that $Z = v^\top A v$. Let A'_{ij} be the symmetric matrix obtained from A by replacing X_{ij} (and X_{ji}) by $x'_{ij} \in [-1,1]$ (and keeping the other entries fixed). Then,

$$\begin{aligned}(Z - Z_{ij})_+ &\le (v^\top A v - v^\top A'_{ij} v)\mathbf{1}_{Z>Z_{ij}} \\ &= \left(v^\top (A - A'_{ij})v\right)\mathbf{1}_{Z>Z_{ij}} \\ &\le 2\left(v_i v_j (X_{ij} - X'_{ij})\right)_+ \le 4|v_i v_j|.\end{aligned}$$

Therefore, using Z_{ij} as defined in (185),

$$\sum_{1\le i\le j\le n} (Z - Z_{ij})^2 = \sum_{1\le i\le j\le n} \inf_{x'_{ij}:|x'_{ij}|\le 1} (Z - Z'_{ij})_+^2 \le \sum_{1\le i\le j\le n} 16|v_i v_j|^2 \le 16\left(\sum_{i=1}^n v_i^2\right)^2 = 16.$$

Thus, by Theorem 13.16 we get

$$\mathbb{P}(Z - \mathbb{E}Z > t) \le e^{-t^2/32}.$$

This example shows that if we want to bound $(Z - Z'_{ij})_+^2$ individually, we get an upper bound of 4 and the usual bounded differences inequality does not lead to a good concentration bound. But this (stronger) version of the bounded difference inequality yields a much stronger result. Note that the above result applies for the adjacency matrix of a random graph if the edges a sampled independently.

13.6 Suprema of the empirical process: exponential inequalities

The location of the distribution of the norm $\|\mathbb{G}_n\|_{\mathcal{F}}$ of the empirical process depends strongly on the complexity of the class of functions $\mathcal{F}$. We have seen various bounds on its the mean value $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}}$ in terms of entropy integrals. It turns out that the spread or "concentration" of the distribution hardly depends on the complexity of the class $\mathcal{F}$. It is sub-Gaussian as soon as the class $\mathcal{F}$ is uniformly bounded, no matter its complexity.

Theorem 13.18. *If $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f(x) - f(y)| \le 1$ for every $f \in \mathcal{F}$ and every $x, y \in \mathcal{X}$, then, for all $t > 0$,*

$$\begin{aligned}\mathbb{P}^*\left(\left|\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}}\right| \ge t\right) &\le 2e^{-2t^2} \qquad \textit{and} \\ \mathbb{P}^*\left(\left|\sup_f \mathbb{G}_n - \mathbb{E}\sup_f \mathbb{G}_n\right| \ge t\right) &\le 2e^{-2t^2}.\end{aligned}$$

This theorem is a special case of the bounded difference inequality; both the norm $\|\mathbb{G}_n\|_{\mathcal{F}}$ and the supremum $\sup_{f\in\mathcal{F}} \mathbb{G}_n$ satisfy the bounded difference property (24) with $c_i n^{-1/2}$ the supremum over f of the range of f (in fact, we can take $c_i = n^{-1/2}$).

References

[Billingsley, 1999] Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition. A Wiley-Interscience Publication.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[Bousquet, 2003] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, Basel.

[Browder, 1996] Browder, A. (1996). *Mathematical analysis*. Undergraduate Texts in Mathematics. Springer-Verlag, New York. An introduction.

[Cantelli, 1933] Cantelli, F. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 4:421–424.

[Chernozhukov et al., 2014] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

[de la Peña and Giné, 1999] de la Peña, V. H. and Giné, E. (1999). *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[Donsker, 1952] Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Komogorov-Smirnov theorems. *Ann. Math. Statistics*, 23:277–281.

[Dudley, 1978] Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929 (1979).

[Dvoretzky et al., 1956] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669.

[Einmahl and Mason, 2005] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403.

[Giné and Nickl, 2016] Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

[Glivenko, 1933] Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 4:92–99.

[Greenshtein and Ritov, 2004] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

[Grenander, 1956] Grenander, U. (1956). On the theory of mortality measurement. I. *Skand. Aktuarietidskr.*, 39:70–96.

[Hjort and Pollard, 2011] Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

[Hoffmann-Jø rgensen, 1991] Hoffmann-Jø rgensen, J. (1991). *Stochastic processes on Polish spaces*, volume 39 of *Various Publications Series (Aarhus)*. Aarhus Universitet, Matematisk Institut, Aarhus.

[Kallenberg, 2002] Kallenberg, O. (2002). *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition.

[Kiefer, 1961] Kiefer, J. (1961). On large deviations of the empiric D. F. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math.*, 11:649–660.

[Kim and Pollard, 1990] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, 18(1):191–219.

[Klein and Rio, 2005] Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077.

[Koltchinskii, 2011] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

[Massart, 1990] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283.

[Pollard, 1984] Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York.

[Pollard, 1989] Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci.*, 4(4):341–366. With comments and a rejoinder by the author.

[Robertson et al., 1988] Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

[Talagrand, 1994] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76.

[Talagrand, 1996a] Talagrand, M. (1996a). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563.

[Talagrand, 1996b] Talagrand, M. (1996b). A new look at independence. *Ann. Probab.*, 24(1):1–34.

[van de Geer, 2000] van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[van der Vaart and Wellner, 2011] van der Vaart, A. and Wellner, J. A. (2011). A local maximal inequality under uniform entropy. *Electron. J. Stat.*, 5:192–203.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[van der Vaart and Wellner, 1996] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. A non-asymptotic viewpoint.