

Uniform laws of large numbers

The focus of this chapter is a class of results known as uniform laws of large numbers. As suggested by their name, these results represent a strengthening of the usual law of large numbers, which applies to a fixed sequence of random variables, to related laws that hold uniformly over collections of random variables. On one hand, such uniform laws are of theoretical interest in their own right, and represent an entrypoint to a rich area of probability and statistics known as empirical process theory. On the other hand, uniform laws also play a key role in more applied settings, including understanding the behavior of different types of statistical estimators. The classical versions of uniform laws are of an asymptotic nature, whereas more recent work in the area has emphasized non-asymptotic results. Consistent with the overall goals of this book, this chapter will follow the non-asymptotic route, presenting results that apply to all sample sizes. In order to do so, we make use of the tail bounds and the notion of Rademacher complexity previously introduced in Chapter 2.

3
4
5
6
7
8
9
10
11
12
13
14
15

■ 4.1 Motivation

16

We begin with some statistical motivations for deriving laws of large numbers, first for the case of cumulative distribution functions (CDFs) and then for more general function classes.

17
18
19

■ 4.1.1 Uniform convergence of CDFs

20

The law of any scalar random variable X can be fully specified by its cumulative distribution function, whose value at any point $t \in \mathbb{R}$ is given by $F(t) := \mathbb{P}[X \leq t]$. Now suppose that we are given a collection $x_1^n = \{x_1, x_2, \dots, x_n\}$ of n i.i.d. samples, each drawn according to the law specified by F . A natural estimate of F is the *empirical CDF* given by

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}[X_i], \quad (4.1)$$

1 where $\mathbb{1}_{(-\infty, t]}[x]$ is a $\{0, 1\}$ -valued indicator function for the event $\{x \leq t\}$. Since the
 2 population CDF can be written as $F(t) = \mathbb{E}[\mathbb{1}_{(-\infty, t]}[X]]$, the empirical CDF is unbiased
 3 in a pointwise sense.

4 Figure 4-1 provides some illustrations of empirical CDFs for the uniform distribution
 5 on the interval $[0, 1]$ for different sample sizes. Note that \widehat{F}_n is a random function, with
 6 the value $\widehat{F}_n(t)$ corresponding to the fraction of samples that lie in the interval $(-\infty, t]$.
 7 Consequently, for any fixed $t \in \mathbb{R}$, the law of large numbers implies that $\widehat{F}_n(t) \xrightarrow{\text{prob.}} F(t)$.

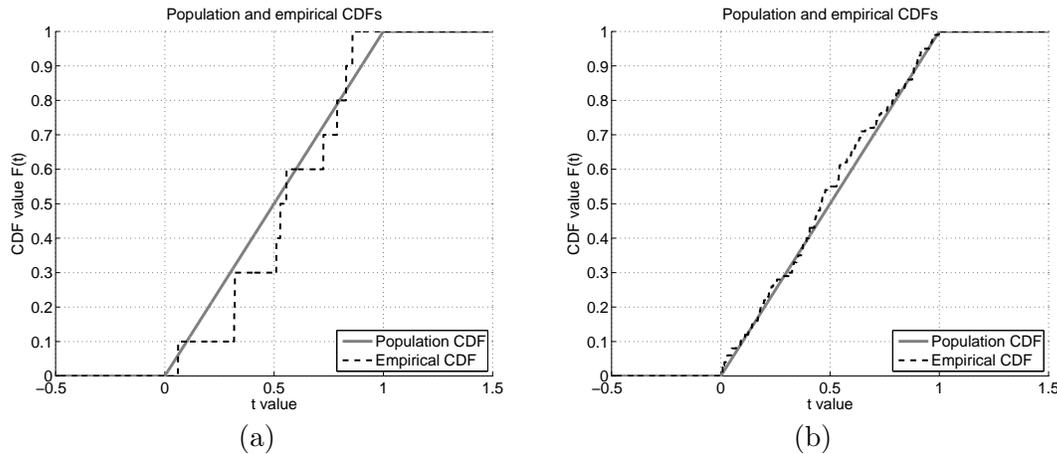


Figure 4-1. Plots of population and empirical CDF functions for the uniform distribution on $[0, 1]$.
 (a) Empirical CDF based on $n = 10$ samples. (b) Empirical CDF based on $n = 100$ samples.

8
 9 In statistical settings, a typical use of the empirical CDF is to construct estimators of
 10 various quantities associated with the population CDF. Many such estimation problems
 11 can be formulated in terms of functional γ that maps any CDF F to a real number
 12 $\gamma(F)$ —that is, $F \mapsto \gamma(F)$. Given a set of samples distributed according to F , the *plug-*
 13 *in principle* suggests replacing the unknown F with the empirical CDF \widehat{F}_n , thereby
 14 obtaining $\gamma(\widehat{F}_n)$ as an estimate of $\gamma(F)$. Let us illustrate this procedure via some
 15 examples.

Example 4.1 (Expectation functionals). Given some integrable function g , we may
 define the *expectation functional* γ_g via

$$\gamma_g(F) := \int g(x) dF(x). \quad (4.2)$$

16 For instance, for the function $g(x) = x$, the functional γ_g maps F to $\mathbb{E}[X]$, where
 17 X is a random variable with CDF F . For any g , the plug-in estimate is given by
 18 $\gamma_g(\widehat{F}_n) = \frac{1}{n} \sum_{i=1}^n g(X_i)$, corresponding to the sample mean of $g(X)$. In the special case

$g(x) = x$, we recover the usual sample mean $\frac{1}{n} \sum_{i=1}^n X_i$ as an estimate for the mean $\mu = \mathbb{E}[X]$. A similar interpretation applies to other choices of the underlying function g . ♣

Example 4.2 (Quantile functionals). For any $\alpha \in [0, 1]$, the *quantile functional* Q_α is given by

$$Q_\alpha(F) := \inf\{t \in \mathbb{R} \mid F(t) \geq \alpha\}. \quad (4.3)$$

The median corresponds to the special case $\alpha = 0.5$. The plug-in estimate is given by

$$Q_\alpha(\widehat{F}_n) := \inf\left\{t \in \mathbb{R} \mid \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[t, \infty)}[X_i] \geq \alpha\right\}, \quad (4.4)$$

and corresponds to estimating the α^{th} quantile of the distribution by the α^{th} sample quantile. In the special case $\alpha = 0.5$, this estimate corresponds to the sample median. Again, it is of interest to determine in what sense (if any) the random variable $Q_\alpha(\widehat{F}_n)$ approaches $Q_\alpha(F)$ as n becomes large. In this case, $Q_\alpha(\widehat{F}_n)$ is a fairly complicated, non-linear function of all the variables, so that this convergence does not follow immediately by a classical result such as the law of large numbers. ♣

Example 4.3 (Goodness-of-fit functionals). It is frequently of interest to test the hypothesis of whether or not a given set of data has been drawn from a known distribution F_0 . For instance, we might be interested in assessing departures from uniformity, in which case F_0 would be a uniform distribution on some interval, or departures from Gaussianity, in which F_0 would specify a Gaussian with a fixed mean and variance. Such tests can be performed using functionals that measure the distance between F and the target CDF F_0 , including the sup-norm distance $\|F - F_0\|_\infty$, or other distances such as the Cramér-von-Mises criterion based on the functional $\gamma(F) := \int_{-\infty}^{\infty} [F(x) - F_0(x)]^2 dF_0(x)$. ♣

For any plug-in estimator $\gamma(\widehat{F}_n)$, an important question is to understand when it is consistent—i.e., when does $\gamma(\widehat{F}_n)$ converges 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 CDFs F and G , let us measure the distance between them using the sup-norm

$$\|G - F\|_\infty := \sup_{t \in \mathbb{R}} |G(t) - F(t)|. \quad (4.5)$$

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 a $\delta > 0$ such that $\|G - F\|_\infty \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \epsilon$.

As we explore in Exercise 4.1, this notion is useful, because for any continuous functional, it reduces the consistency question for the plug-in estimator $\gamma(\widehat{F}_n)$ to the issue of whether or not $\|\widehat{F}_n - F\|_\infty$ converges to zero. A classical result, known as the Glivenko-Cantelli theorem, addresses the latter question:

Theorem 4.1 (Glivenko-Cantelli). For any distribution, the empirical CDF \widehat{F}_n is a strongly consistent estimator of the population CDF F in the uniform norm, meaning that

$$\|\widehat{F}_n - F\|_\infty \xrightarrow{a.s.} 0. \quad (4.6)$$

We provide a proof of this claim as a corollary of a more general result to follow (see Theorem 4.2). For statistical applications, an important consequence of Theorem 4.1 is that the plug-in estimate $\gamma(\widehat{F}_n)$ is almost surely consistent as an estimator of $\gamma(F)$ for any functional γ that is continuous with respect to the sup-norm. See Exercise 4.1 for further exploration of these issues.

4.1.2 Uniform laws for more general function classes

We now turn to more general consideration of uniform laws of large numbers. Let \mathcal{F} be a class of integrable real-valued functions with domain \mathcal{X} , and let $X_1^n = \{X_1, \dots, X_n\}$ be a collection of i.i.d. samples from some distribution \mathbb{P} over \mathcal{X} . Consider the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f] \right|, \quad (4.7)$$

which measures the absolute deviation between the sample average $\frac{1}{n} \sum_{i=1}^n f(X_i)$ and the population average $\mathbb{E}[f] = \mathbb{E}[f(X)]$, uniformly over the class \mathcal{F} . (See the bibliographic section for a discussion of possible measurability concerns with this random variable.)

Definition 4.1. We say that \mathcal{F} is a *Glivenko-Cantelli* class for \mathbb{P} if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ converges to zero in probability as $n \rightarrow \infty$.

This notion can also be defined in a stronger sense, requiring almost sure convergence of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, in which case we say that \mathcal{F} satisfies a *strong Glivenko-Cantelli law*. The classical result on the empirical CDF (Theorem 4.1) can be reformulated as a particular case of this notion:

Example 4.4 (Empirical CDFs and indicator functions). Consider the function class

$$\mathcal{F} = \{\mathbb{1}_{(-\infty, t]}(\cdot) \mid t \in \mathbb{R}\}, \quad (4.8)$$

where $\mathbb{1}_{(-\infty, t]}$ is the $\{0, 1\}$ -valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have $\mathbb{E}[\mathbb{1}_{(-\infty, t]}(X)] = \mathbb{P}[X \leq t] = F(t)$, so that the classical Glivenko-Cantelli theorem corresponds to a strong uniform law for the class (4.8). ♣

Not all classes of functions are Glivenko-Cantelli, as illustrated by the following example.

Example 4.5 (Failure of uniform law). Let \mathcal{S} be the class of all subsets S of $[0, 1]$ such that the subset S has a finite number of elements, and consider the function class $\mathcal{F}_{\mathcal{S}} = \{\mathbb{1}_S(\cdot) \mid S \in \mathcal{S}\}$ of ($\{0-1\}$ -valued) indicator functions of such sets. Suppose that samples X_i are drawn from some distribution over $[0, 1]$ that has no atoms (i.e., $\mathbb{P}(\{x\}) = 0$ for all $x \in [0, 1]$); this class includes any distribution that has a density with respect to Lebesgue measure. For any such distribution, we are guaranteed that $\mathbb{P}[S] = 0$ for all $S \in \mathcal{S}$. On the other hand, for any positive integer $n \in \mathbb{N}$, the set $X_1^n = \{X_1, \dots, X_n\}$ belongs to \mathcal{S} , and moreover by definition of the empirical distribution, we have $\mathbb{P}_n[X_1^n] = 1$. Putting together the pieces, we conclude that

$$\sup_{S \in \mathcal{S}} |\mathbb{P}_n[S] - \mathbb{P}[S]| = 1 - 0 = 1, \quad (4.9)$$

so that the function class $\mathcal{F}_{\mathcal{S}}$ is *not* a Glivenko-Cantelli class for \mathbb{P} . ♣

We have seen that the classical Glivenko-Cantelli law—which guarantees convergence of a special case of the variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ —is of interest in analyzing estimators based on “plugging in” the empirical CDF. It is natural to ask in what other statistical contexts do these quantities arise? In fact, variables of the form $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ are ubiquitous throughout statistics—in particular, they lie at the heart of methods based on empirical risk minimization. In order to describe this notion more concretely, let us consider an indexed-family of probability distributions $\{\mathbb{P}_{\theta} \mid \theta \in \Omega\}$, and suppose that we are given n samples $X_1^n = \{X_1, \dots, X_n\}$, each sample lying in some space \mathcal{X} . Suppose that the samples are drawn i.i.d. according to a distribution \mathbb{P}_{θ^*} , for some fixed but unknown $\theta^* \in \Omega$. Here the index θ^* could lie within a finite-dimensional space, such as $\Omega = \mathbb{R}^d$ in a vector estimation problem, or could lie within some function class $\Omega = \mathcal{G}$, in which case the problem is of the non-parametric variety.

In either case, a standard decision-theoretic approach to estimating θ^* is based on minimizing a loss function of the form $\mathcal{L}_{\theta}(x)$, which measures the “fit” between a parameter $\theta \in \Omega$ and the sample $X \in \mathcal{X}$. Given the collection of n samples X_1^n , the

principle of empirical risk minimization is based on the objective function

$$\widehat{R}_n(\theta, \theta^*) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\theta(X_i).$$

This quantity is known as the *empirical risk*, since it is defined by the samples X_1^n , and our notation reflects the fact that these samples depend—in turn—on the unknown distribution \mathbb{P}_{θ^*} . This empirical risk should be contrasted with the *population risk*

$$R(\theta, \theta^*) := \mathbb{E}_{\theta^*}[\mathcal{L}_\theta(X)],$$

- 1 where the expectation \mathbb{E}_{θ^*} is taken over a sample $X \sim \mathbb{P}_{\theta^*}$.

In practice, one minimizes the empirical risk over some subset Ω_0 of the full space Ω , thereby obtaining some estimate $\widehat{\theta}$. The statistical question is how to bound the *excess risk*, measured in terms of the population quantities—namely the difference

$$\delta R(\widehat{\theta}, \theta^*) := R(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega_0} R(\theta, \theta^*).$$

- 2 Let us consider some examples to illustrate.

Example 4.6 (Maximum likelihood). Consider a family of distributions $\{\mathbb{P}_\theta, \theta \in \Omega\}$, each with a strictly positive density p_θ (defined with respect to a common underlying measure). Now suppose that we are given n i.i.d. samples from unknown distribution \mathbb{P}_{θ^*} , and we would like to estimate the unknown parameter θ^* . In order to do so, we consider the cost function

$$\mathcal{L}_\theta(x) := \log \frac{p_{\theta^*}(x)}{p_\theta(x)}.$$

(The term $p_{\theta^*}(x)$, which we have included for later theoretical convenience, has no effect on the minimization over θ .) Indeed, the maximum likelihood estimate is obtained by minimizing the empirical risk

$$\widehat{\theta} \in \arg \min_{\theta \in \Omega_0} \underbrace{\frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}}_{\widehat{R}_n(\theta, \theta^*)} = \arg \min_{\theta \in \Omega_0} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{p_\theta(X_i)}$$

- 3 The population risk is given by $R(\theta, \theta^*) = \mathbb{E}_{\theta^*}[\log \frac{p_{\theta^*}(X)}{p_\theta(X)}]$, a quantity known as the
 4 *Kullback-Leibler divergence* between p_{θ^*} and p_θ . In the special case that $\theta^* \in \Omega_0$, the
 5 excess risk is simply the Kullback-Leibler divergence between the true density p_{θ^*} and
 6 the fitted model $p_{\widehat{\theta}}$. 

Example 4.7 (Binary classification). Suppose that we observe n samples of the form $(X_i, Y_i) \in \{-1, +1\} \times \mathbb{R}^d$, where the vector X_i corresponds to a set of d predictors or features, and the binary variable Y_i corresponds to a label. We can view such data as being generated by some distribution \mathbb{P}_X over the features, and a conditional distribution $\mathbb{P}_{Y|Z}$. Since Y takes binary values, the conditional distribution is fully specified by the likelihood ratio $\psi(x) = \frac{\mathbb{P}[Y=+1|X=x]}{\mathbb{P}[Y=-1|X=x]}$.

The goal of binary classification is to estimate a function $f : \mathbb{R}^d \rightarrow \{-1, +1\}$ that minimizes the probability of mis-classification $\mathbb{P}[f(X) \neq Y]$, for an independently drawn pair (X, Y) . Note that this probability of error corresponds to the population risk for the cost function

$$\mathcal{L}_f(X, Y) := \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{otherwise.} \end{cases} \quad (4.10)$$

A function that minimizes this probability of error is known as a *Bayes classifier* f^* ; in the special case of equally probable classes ($\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$), a Bayes classifier is given by $f^*(x) = +1$ if $\phi(x) \geq 1$, and $f^*(x) = -1$ otherwise. Since the likelihood ratio ϕ (and hence f^*) is unknown, a natural approach to approximating the Bayes rule is based on choosing \hat{f} to minimize the empirical risk

$$\hat{R}_n(f, f^*) := \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}[f(X_i) \neq Y_i]}_{\mathcal{L}_f(X_i, Y_i)},$$

corresponding to the fraction of training samples that are mis-classified. Typically, the minimization over f is restricted to some subset of all possible decision rules. See Chapter 14 for some further discussion of how to analyze such methods for binary classification. 

Returning to the main thread, our goal is to develop methods for controlling the excess risk. For simplicity, let us assume¹ that there exists some $\theta_0 \in \Omega_0$ such that $R(\theta_0, \theta^*) = \inf_{\theta \in \Omega_0} R(\theta, \theta^*)$. With this notation, the excess risk can be decomposed as

$$\delta R(\hat{\theta}, \theta^*) = \underbrace{\{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)\}}_{T_1} + \underbrace{\{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)\}}_{T_2 \leq 0} + \underbrace{\{\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)\}}_{T_3}.$$

Note that the middle term is non-positive, since $\hat{\theta}$ minimizes the empirical risk over Ω_0 .

The third term can be dealt with in a relatively straightforward manner, because θ_0 is an unknown but non-random quantity. Indeed, recalling the definition of the

¹If the infimum is not achieved, then we choose an element θ_0 for which this equality holds up to some arbitrarily small tolerance $\epsilon > 0$, and the analysis to follow holds up to this tolerance.

empirical risk, we have

$$T_3 = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\theta_0}(X_i) - \mathbb{E}[\mathcal{L}_{\theta_0}(X)],$$

corresponding to the deviation of a sample mean from its expectation for the random variable $\mathcal{L}_{\theta_0}(X)$. This quantity can be controlled using the techniques introduced in Chapter 2—for instance, via the Hoeffding bound when the samples are independent and the loss function is bounded. The first term can be written in a similar way, namely as the sum

$$T_1 = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\hat{\theta}}(X_i) - \mathbb{E}[\mathcal{L}_{\hat{\theta}}(X)].$$

This quantity is more challenging to control, because the parameter $\hat{\theta}$ —in sharp contrast to the deterministic quantity θ_0 —is now random. In general, it depends on all the samples X_1^n , since it was obtained by minimizing the empirical risk. For this reason, controlling the first term requires a stronger result, such as a uniform law of large numbers over the loss class $\mathfrak{L}(\Omega_0) := \{\mathcal{L}_\theta, \theta \in \Omega_0\}$. With this notation, we have

$$T_1 \leq \sup_{\theta \in \Omega_0} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\theta(X_i) - \mathbb{E}[\mathcal{L}_\theta(X)] \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{L}(\Omega_0)}$$

- 1 Since T_3 is also dominated by this same quantity, we conclude that the excess risk is
- 2 at most $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{L}(\Omega_0)}$. This derivation demonstrates that the central challenge in
- 3 analyzing estimators based on empirical risk minimization is to establish a uniform law
- 4 of large numbers for the loss class $\mathfrak{L}(\Omega_0)$. We explore various concrete examples of this
- 5 procedure in the exercises.

6 ■ 4.2 A uniform law via Rademacher complexity

Having developed various motivations for studying uniform laws, let us now turn to the technical details of deriving such results. An important quantity that underlies the study of uniform laws is the *Rademacher complexity* of the function class \mathcal{F} . For any fixed collection $x_1^n := (x_1, \dots, x_n)$ of points, consider the subset of \mathbb{R}^n given by

$$\mathcal{F}(x_1^n) := \{(f(x_1), \dots, f(x_n)) \mid f \in \mathcal{F}\}. \quad (4.11)$$

The set $\mathcal{F}(x_1^n)$ corresponds to all those vectors in \mathbb{R}^n that can be realized by applying a function $f \in \mathcal{F}$ to the collection (x_1, \dots, x_n) , and the *empirical Rademacher complexity*

is given by

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \quad (4.12)$$

Note that this definition coincides with our earlier definition of the Rademacher complexity of a set (see Example 2.11).

Given a collection $X_1^n := (X_1, \dots, X_n)$ of random samples, then the empirical Rademacher complexity $\mathcal{R}(\mathcal{F}(X_1^n)/n)$ is a random variable. Taking its expectation yields the *Rademacher complexity of the function class* \mathcal{F} —namely, the deterministic quantity

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_X [\mathcal{R}(\mathcal{F}(X_1^n)/n)] = \mathbb{E}_{X, \varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]. \quad (4.13)$$

Note that the Rademacher complexity is the average of the maximum correlation between the vector $(f(X_1), \dots, f(X_n))$ and the “noise vector” $(\varepsilon_1, \dots, \varepsilon_n)$, where the maximum is taken over all functions $f \in \mathcal{F}$. The intuition is a natural one: a function class is extremely large—and in fact, “too large” for statistical purposes—if we can always find a function that has a high correlation with a randomly drawn noise vector. Conversely, when the Rademacher complexity decays as a function of sample size, then it is impossible to find a function that correlates very highly in expectation with a randomly drawn noise vector.

We now make precise the connection between Rademacher complexity and the Glivenko-Cantelli property, in particular by showing that any bounded function class \mathcal{F} with $\mathcal{R}_n(\mathcal{F}) = o(1)$ is also a Glivenko-Cantelli class. More precisely, we prove a non-asymptotic statement, in terms of a tail bound for the probability that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ deviates substantially above a multiple of the Rademacher complexity.

Theorem 4.2. Let \mathcal{F} be a class of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ that is uniformly bounded (i.e., $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$). Then for all $n \geq 1$ and $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta \quad (4.14)$$

with \mathbb{P} -probability at least $1 - 2 \exp(-\frac{n\delta^2}{8b^2})$. Consequently, as long as $\mathcal{R}_n(\mathcal{F}) = o(1)$, we have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$.

In order for Theorem 4.2 to be useful, we need to obtain upper bounds on the

1 Rademacher complexity. There are a variety of methods for doing so, ranging from
 2 direct calculations to alternative complexity measures. In Section 4.3, we develop some
 3 techniques for upper bounding the Rademacher complexity for indicator functions of
 4 half-intervals, as required for the classical Glivenko-Cantelli theorem (see Example 4.4);
 5 we also discuss the notion of Vapnik-Chervonenkis dimension, which can be used to up-
 6 per bound the Rademacher complexity for other function classes. In Chapter 5, we
 7 introduce more advanced techniques based on metric entropy and chaining for control-
 8 ling Rademacher complexity and related sub-Gaussian processes. In the meantime, let
 9 us turn to the proof of Theorem 4.2.

10 *Proof.* We first note if $\mathcal{R}_n(\mathcal{F}) = o(1)$, then the almost-sure convergence follows from
 11 the tail bound (4.14) and the Borel-Cantelli lemma. Accordingly, the remainder of the
 12 argument is devoted to proving the tail bound (4.14).

Concentration around mean: We first claim that when \mathcal{F} is uniformly bounded, then the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is sharply concentrated around its mean. In order to simplify notation, it is convenient to define the re-centered functions $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, and to write $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(X_i) \right|$. Thinking of the samples as fixed for the moment, consider the function $G(x_1, \dots, x_n) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|$. We claim that G satisfies the Lipschitz property required to apply the bounded differences method (recall Corollary 2.2). Since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first co-ordinate x_1 is perturbed. Accordingly, we define the vector $y \in \mathbb{R}^n$ with $y_i = x_i$ for all $i \neq 1$, and seek the bound the difference $|G(x) - G(y)|$. For any function $\bar{f} = f - \mathbb{E}[f]$, we have

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \sup_{h \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{h}(y_i) \right| &\leq \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| \\ &\leq \frac{1}{n} |\bar{f}(x_1) - \bar{f}(y_1)| \\ &\leq \frac{2b}{n}, \end{aligned} \tag{4.15}$$

where the final inequality uses the fact that

$$|\bar{f}(x_1) - \bar{f}(y_1)| = |f(x_1) - f(y_1)| \leq 2b,$$

which follows from the uniform boundedness condition $\|f\|_{\infty} \leq b$. Since the inequality (4.15) holds for any function f , we may take the supremum over $f \in \mathcal{F}$ on both sides; doing so yields the inequality $G(x) - G(y) \leq \frac{2b}{n}$. Since the same argument may be applied with the roles of x and y reversed, we conclude that $|G(x) - G(y)| \leq \frac{2b}{n}$.

Therefore, by the bounded differences method (see Corollary 2.2), we have

$$\left| \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \right| \leq t \quad \text{with } \mathbb{P}\text{-prob. at least } 1 - 2 \exp\left(-\frac{nt^2}{8b^2}\right), \quad (4.16)$$

valid for all $t \geq 0$.

Upper bound on mean: It remains to show that $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}]$ is upper bounded by at most $2\mathcal{R}_n(\mathcal{F})$, and we do so using a classical symmetrization argument. Letting (Y_1, \dots, Y_n) be a second i.i.d. sequence, independent of (X_1, \dots, X_n) , we have

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

where the inequality follows from Jensen's inequality for convex functions. Now let $(\varepsilon_1, \dots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y . For any function $f \in \mathcal{F}$ and any $i = 1, 2, \dots, n$, the variable $\varepsilon_i(f(X_i) - f(Y_i))$ has the same distribution as $f(X_i) - f(Y_i)$, whence

$$\begin{aligned} \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &\leq \mathbb{E}_{X,Y,\varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(Y_i)) \right| \right] \\ &\leq 2 \mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] = 2 \mathcal{R}_n(\mathcal{F}). \quad (4.17) \end{aligned}$$

Combining the upper bound (4.17) with the tail bound (4.16) yields the claim. \square

■ 4.2.1 Necessary conditions with Rademacher complexity

The proof of Theorem 4.2 illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to its symmetrized version

$$\|R_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \quad (4.18)$$

Note that the expectation of $\|R_n\|_{\mathcal{F}}$ corresponds to the Rademacher complexity, which plays a central role in Theorem 4.2. It is natural to wonder whether much was lost in moving from the variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to its symmetrized version. The following “sandwich” result relates these quantities.

Proposition 4.1. For any convex non-decreasing function $\Phi : \mathbb{R} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_{X,\varepsilon}[\Phi(\frac{1}{2}\|R_n\|_{\overline{\mathcal{F}}})] \stackrel{(a)}{\leq} \mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \stackrel{(b)}{\leq} \mathbb{E}_{X,\varepsilon}[\Phi(2\|R_n\|_{\mathcal{F}})], \quad (4.19)$$

where $\overline{\mathcal{F}} = \{f - \mathbb{E}[f], f \in \mathcal{F}\}$ is the re-centered function class.

When applied with the convex non-decreasing function $\Phi(t) = t$, Proposition 4.1 yields the inequalities

$$\frac{1}{2}\mathbb{E}_{X,\varepsilon}\|R_n\|_{\overline{\mathcal{F}}} \leq \mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq 2\mathbb{E}_{X,\varepsilon}\|R_n\|_{\mathcal{F}}, \quad (4.20)$$

with the only difference being the use of \mathcal{F} in the upper bound, and the re-centered class $\overline{\mathcal{F}}$ in the lower bound.

Other choices of interest include $\Phi(t) = e^{\lambda t}$ for some $\lambda > 0$, which can be used to control the moment generating function.

Proof. The proof of inequality (b) is essentially identical to the argument for $\Phi(t) = t$ provided in the proof of Theorem 4.2. Turning to the bound (a), we have

$$\begin{aligned} \mathbb{E}_{X,\varepsilon}[\Phi(\frac{1}{2}\|R_n\|_{\overline{\mathcal{F}}})] &= \mathbb{E}_{X,\varepsilon}[\Phi(\frac{1}{2}\sup_{f \in \overline{\mathcal{F}}} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i \{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}|)] \\ &\stackrel{(i)}{\leq} \mathbb{E}_{X,Y,\varepsilon}[\Phi(\frac{1}{2}\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i \{f(X_i) - f(Y_i)\}|)] \\ &\stackrel{(ii)}{=} \mathbb{E}_{X,Y}[\Phi(\frac{1}{2}\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(Y_i)\}|)] \end{aligned}$$

where inequality (i) follows from Jensen's inequality and the convexity of Φ ; and equality (ii) follows since for each $i = 1, 2, \dots, n$ and $f \in \mathcal{F}$, the variables $\varepsilon_i \{f(X_i) - f(Y_i)\}$ and $f(X_i) - f(Y_i)$ have the same distribution. Adding and subtracting $\mathbb{E}[f]$ and applying triangle inequality, we obtain that $T := \frac{1}{2}\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(Y_i)\}|$ is upper bounded as

$$T \leq \frac{1}{2}\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}[f]\}| + \frac{1}{2}\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(Y_i) - \mathbb{E}[f]\}|$$

Since Φ is convex and non-decreasing, we obtain

$$\Phi(T) \leq \frac{1}{2}\Phi(\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}[f]\}|) + \frac{1}{2}\Phi(\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \{f(Y_i) - \mathbb{E}[f]\}|).$$

Since X and Y are identically distributed, taking expectations yields the claim. \square

A consequence of Proposition 4.1 is that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ can be lower bounded by a multiple of Rademacher complexity, and some fluctuation terms. This fact can be used to prove the following:

Proposition 4.2. Let \mathcal{F} be a class of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ that is uniformly bounded (i.e., $\|f\|_{\infty} \leq b$ for all $f \in \mathcal{F}$). Then for all $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \frac{1}{2} \mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f \in \mathcal{F}} |\mathbb{E}[f]|}{2\sqrt{n}} - \delta \quad (4.21)$$

with \mathbb{P} -probability at least $1 - 2e^{-\frac{n\delta^2}{8b^2}}$.

We leave the proof of this result for the reader (see Exercise 4.3). As a consequence, if the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ remains bounded away from zero, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ cannot converge to zero in probability. We have thus shown that for a uniformly bounded function class \mathcal{F} , the Rademacher complexity provides a necessary and sufficient condition for it to be Glivenko-Cantelli.

■ 4.3 Upper bounds on the Rademacher complexity

Obtaining concrete results using Theorem 4.2 requires methods for upper bounding the Rademacher complexity. There are a variety of such methods, ranging from simple union bound methods (suitable for finite function classes) to more advanced techniques involving the notion of metric entropy and chaining arguments. We explore the latter techniques in Chapter 5 to follow. This section is devoted to more elementary techniques, including those required to prove the classical Glivenko-Cantelli result, and more generally, those that apply to function classes with polynomial discrimination.

■ 4.3.1 Classes with polynomial discrimination

For a given collection of points $x_1^n = (x_1, \dots, x_n)$, the “size” of the set $\mathcal{F}(x_1^n)$ provides a sample-dependent measure of the complexity of \mathcal{F} . In the simplest case, the set $\mathcal{F}(x_1^n)$ contains only a finite number of vectors for all sample sizes, so that its “size” can be measured via its cardinality. For instance, if \mathcal{F} consists of a family of decision rules taking binary values (as in Example 4.7), then $\mathcal{F}(x_1^n)$ can contain at most 2^n elements. Of interest to us are function classes for which this cardinality grows only as a polynomial function of n , as formalized in the following:

Definition 4.2 (Polynomial discrimination). A class \mathcal{F} of functions with domain \mathcal{X} has polynomial discrimination of order $\nu \geq 1$ if for each positive integer n and collection $x_1^n = \{x_1, \dots, x_n\}$ of n points in \mathcal{X} , the set $\mathcal{F}(x_1^n)$ has cardinality upper bounded as

$$\text{card}(\mathcal{F}(x_1^n)) \leq (n+1)^\nu. \quad (4.22)$$

The significance of this property is that it provides a straightforward approach to controlling the Rademacher complexity. For any set $S \subset \mathbb{R}^n$, we let $D := \sup_{x \in S} \|x\|_2$ denote its maximal width in the ℓ_2 -norm.

Lemma 4.1. Suppose that \mathcal{F} has polynomial discrimination of order ν . Then for all $n \geq 10$ and any collection of points $x_1^n = (x_1, \dots, x_n)$,

$$\mathbb{E}_\varepsilon \left[\underbrace{\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right|}_{\mathcal{R}(\mathcal{F}(x_1^n)/n)} \right] \leq 3 \sqrt{\frac{D^2(x_1^n) \nu \log(n+1)}{n}},$$

where $D(x_1^n) := \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2 -radius of the set $\mathcal{F}(x_1^n)/\sqrt{n}$.

We leave the proof of this claim for the reader (see Exercise 4.7).

Although Lemma 4.1 is stated as an upper bound on the empirical Rademacher complexity, it yields as a corollary an upper bound on the Rademacher complexity $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X[\mathcal{R}(\mathcal{F}(X_1^n/n))]$, one which involves the expected ℓ_2 -width $\mathbb{E}_{X_1^n}[D(X)]$. An especially simple case is when function class is uniformly bounded—say $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$ —so that $D(x_1^n) \leq b$ for all samples, and hence Lemma 4.1 implies that

$$\mathcal{R}_n(\mathcal{F}) \leq 3 \sqrt{\frac{b^2 \nu \log(n+1)}{n}} \quad \text{for all } n \geq 10. \quad (4.23)$$

Combined with Theorem 4.2, we conclude that any bounded function class with polynomial discrimination is Glivenko-Cantelli.

What types of function classes have polynomial discrimination? As discussed previously in Example 4.4, the classical Glivenko-Cantelli law is based on indicator functions of the left-sided intervals $(-\infty, t]$. These functions are uniformly bounded with $b = 1$, and moreover, as shown in the following proof, this function class has polynomial discrimination of order $d = 1$. Consequently, Theorem 4.2 combined with Lemma 4.1 yields a quantitative version of Theorem 4.1 as a corollary.

Corollary 4.1 (Classical Glivenko-Cantelli). Let $F(t) = \mathbb{P}[X \geq t]$ be the CDF of a random variable X , and let \widehat{F}_n be the empirical CDF based on n i.i.d. samples $X_i \sim \mathbb{P}$. Then

$$\mathbb{P} \left[\|\widehat{F}_n - F\|_\infty \geq 3 \sqrt{\frac{\log(n+1)}{n}} + \delta \right] \leq 2e^{-\frac{n\delta^2}{8}} \quad \text{for all } \delta \geq 0, \quad (4.24)$$

and hence $\|\widehat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$.

Proof. For a given sample $x_1^n = (x_1, \dots, x_n) \in \mathbb{R}^n$, consider the set $\mathcal{F}(x_1^n)$, where \mathcal{F} is the set of all $\{0-1\}$ -valued indicator functions of the half-intervals $[t, \infty)$ for $t \in \mathbb{R}$. If we order the samples as $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, then they split the real line into at most $n+1$ intervals (including the two end-intervals $(-\infty, x_{(1)})$ and $[x_{(n)}, \infty)$). For a given t , the indicator function $\mathbb{1}_{[t, \infty)}$ takes the value one for all $x_{(i)} \geq t$, and the value zero for all other samples. Thus, we have shown that for any given sample x_1^n , we have $\text{card}(\mathcal{F}(x_1^n)) \leq n+1$. Applying Lemma 4.1, we obtain

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] \leq 3 \sqrt{\frac{\log(n+1)}{n}},$$

and taking averages over the data X_i yields the upper bound $\mathcal{R}_n(\mathcal{F}) \leq 3 \sqrt{\frac{\log(n+1)}{n}}$. The claim (4.24) then follows from Theorem 4.2. \square

Although the exponential tail bound (4.24) is adequate for many purposes, it is far from the tightest possible. Using alternative methods, we provide a sharper result that removes the $\sqrt{\log(n+1)}$ factor in Chapter 5. See the bibliographic section for references to the sharpest possible results, including control of the constants in the exponent and the pre-factor.

■ 4.3.2 Vapnik-Chervonenkis dimension

Thus far, we have seen that it is relatively straightforward to establish uniform laws for function classes with polynomial discrimination. In certain cases, such as in our proof of the classical Glivenko-Cantelli law, we can verify by direct calculation that a given function class has polynomial discrimination. More broadly, it is of interest to develop techniques for certifying this property, and the theory of Vapnik-Chervonenkis (VC) dimension provides one such class of techniques.

Let us consider a function class \mathcal{F} in which each function f is binary-valued, taking the values $\{0, 1\}$ for concreteness. In this case, the set $\mathcal{F}(x_1^n)$ from equation (4.11) can have at most 2^n elements.

Definition 4.3 (Shattering and VC dimension). Given a class \mathcal{F} of binary valued functions, the set $x_1^n = (x_1, \dots, x_n)$ is *shattered* by \mathcal{F} means that $\text{card}(\mathcal{F}(x_1^n)) = 2^n$. The *VC-dimension* $\nu(\mathcal{F})$ is the largest integer n for which there is *some* collection $x_1^n = (x_1, \dots, x_n)$ of n points that can be shattered by \mathcal{F} .

When the quantity $\nu(\mathcal{F})$ is finite, then the function class \mathcal{F} is said to be a *VC-class*. We will frequently consider function classes \mathcal{F} that consist of indicator functions $\mathbb{1}_S[\cdot]$, for sets S ranging over some class of sets \mathcal{S} . In this case, we use $\mathcal{S}(x_1^n)$ and $\nu(\mathcal{S})$ as shorthands for the sets $\mathcal{F}(x_1^n)$ and the VC dimension of \mathcal{F} , respectively.

Let us illustrate the notions of shattering and VC dimension with some examples:

Example 4.8 (Intervals in \mathbb{R}). Consider the class of all indicator functions for left-sided half-intervals on the real line—namely, the class $\mathcal{S}_{\text{left}} := \{(-\infty, a] \mid a \in \mathbb{R}\}$. Implicit in the proof of Corollary 4.1 is a calculation of the VC dimension for this class. We first note that for any single point x_1 , both subsets ($\{x_1\}$ and the empty set \emptyset) can be picked out by the class of left-sided intervals $\{(-\infty, a] \mid a \in \mathbb{R}\}$. But given two distinct points $x_1 < x_2$, it is impossible to find a left-sided interval that contains x_2 but not x_1 . Therefore, we conclude that $\nu(\mathcal{S}_{\text{left}}) = 1$. In the proof of Corollary 4.1, we showed more specifically that for any collection $x_1^n = \{x_1, \dots, x_n\}$, we have $\text{card}(\mathcal{S}_{\text{left}}(x_1^n)) \leq n + 1$.

Now consider the class of all two-sided intervals over the real line—namely, the class $\mathcal{S}_{\text{two}} := \{(b, a] \mid a, b \in \mathbb{R} \text{ such that } b < a\}$. The class \mathcal{S}_{two} can shatter any two-point set. However, given three distinct points $x_1 < x_2 < x_3$, it cannot pick out the subset $\{x_1, x_3\}$, showing that $\nu(\mathcal{S}_{\text{two}}) = 2$. For future reference, let us also upper bound the shattering coefficient of \mathcal{S}_{two} . Note that any collection of n distinct points $x_1 < x_2 < \dots < x_{n-1} < x_n$ divides up the real line into $(n + 1)$ intervals. Thus, any set of the form $(-b, a]$ can be specified by choosing one of $(n + 1)$ intervals for b , and a second interval for a . Thus, a crude upper bound on the shatter coefficient is

$$\text{card}(\mathcal{S}_{\text{two}}(x_n)) \leq (n + 1)^2,$$

showing that this class has polynomial discrimination with degree $\nu = 2$. ♣

Thus far, we have seen two examples of function classes with finite VC dimension, both of which turned out to also polynomial discrimination. Is there a general connection between the VC dimension and polynomial discriminability? Indeed, it turns out that any finite VC class has polynomial discrimination with degree at most the VC dimension; this fact is a deep result that was proved independently (in slightly different forms) by Vapnik and Chervonenkis, Sauer and Shelah. We refer the reader to the bibliographic section for further discussion and references.

In order to understand why this fact is surprising, note that for a given set class \mathcal{S} ,

the definition of VC dimension implies that for all $n > \nu(\mathbb{S})$, we must have $\text{card}(\mathbb{S}(x_1^n)) < 2^n$ for all collections x_1^n of n samples. However, at least in principle, there could exist some subset with

$$\text{card}(\mathbb{S}(x_1^n)) = 2^n - 1,$$

which is not significantly different than 2^n . The following result shows that this is *not* the case; indeed, for any VC-class, the cardinality of $\mathbb{S}(x_1^n)$ can grow at most polynomially in n .

Proposition 4.3 (Vapnik-Chervonenkis, Sauer and Shelah). Consider a set class \mathbb{S} with $\nu(\mathbb{S}) < \infty$. Then for any collection of points $x_1^n = (x_1, \dots, x_n)$, we have

$$\text{card}(\mathbb{S}(x_1^n)) \stackrel{(i)}{\leq} \sum_{i=0}^{\nu(\mathbb{S})} \binom{n}{i} \stackrel{(ii)}{\leq} (n+1)^{\nu(\mathbb{S})}. \quad (4.25)$$

We prove inequality (i) in the Appendix. Given inequality (i), inequality (ii) can be established by elementary combinatorial arguments, so we leave it as an exercise for the reader (part (a) of Exercise 4.11). Part (b) of the same exercise establishes a sharper upper bound.

■ 4.3.3 Controlling the VC dimension

Since classes with finite VC dimension have polynomial discrimination, it is of interest to develop techniques for controlling the VC dimension.

Basic operations

The property of having finite VC dimension is preserved under a number of basic operations, as summarized in the following.

Proposition 4.4. Let \mathbb{S} and \mathbb{T} be set classes, each with finite VC dimensions $\nu(\mathbb{S})$ and $\nu(\mathbb{T})$ respectively. Then each of the following set classes also have finite VC dimension:

- (a) The set class $\mathbb{S}^c := \{S^c \mid S \in \mathbb{S}\}$, where S^c denotes the complement of S .
- (b) The set class $\mathbb{S} \sqcup \mathbb{T} := \{S \cup T \mid S \in \mathbb{S}, T \in \mathbb{T}\}$.
- (c) The set class $\mathbb{S} \cap \mathbb{T} := \{S \cap T \mid S \in \mathbb{S}, T \in \mathbb{T}\}$.

We leave the proof of this result as an exercise for the reader.

1 **Vector space structure**

2 Any class \mathcal{G} of real-valued functions defines a class of sets by the operation of taking
 3 subgraphs. In particular, given a real-valued function $g : \mathcal{X} \rightarrow \mathbb{R}$, its subgraph at level
 4 zero is the subset $S_g := \{x \in \mathcal{X} \mid g(x) \leq 0\}$. In this way, we can associate to \mathcal{G} the
 5 collection of subsets $\mathfrak{S}(\mathcal{G}) := \{S_g, g \in \mathcal{G}\}$, which we refer to as the subgraph class
 6 of \mathcal{G} . Many interesting classes of sets are naturally defined in this way, among them
 7 half-spaces, ellipsoids and so on. In many cases, the underlying function class \mathcal{G} is a
 8 vector space, and the following result allows us to upper bound the VC dimension of
 9 the associated set class $\mathfrak{S}(\mathcal{G})$.

10 **Proposition 4.5** (Finite-dimensional vector spaces). Let \mathcal{G} be a vector space of
 11 functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ with dimension $\dim(\mathcal{G}) < \infty$. Then the subgraph class $\mathfrak{S}(\mathcal{G})$
 12 has VC dimension at most $\dim(\mathcal{G})$.

Proof. By the definition of VC dimension, we need to show that no collection of
 $n = \dim(\mathcal{G}) + 1$ points in \mathbb{R}^d can be shattered by $\mathfrak{S}(\mathcal{G})$. Fix a collection $x_1^n = \{x_1, \dots, x_n\}$
 of n points in \mathbb{R}^d , and consider the linear map $L : \mathcal{G} \rightarrow \mathbb{R}^n$ given by $L(g) = (g(x_1), \dots, g(x_n))$.
 By construction, the range of the mapping L is a linear subspace of \mathbb{R}^n with dimension
 at most $\dim(\mathcal{G}) = n - 1 < n$. Therefore, there must exist a non-zero vector $\gamma \in \mathbb{R}^n$ such
 that $\langle \gamma, L(g) \rangle = 0$ for all $g \in \mathcal{G}$. We may assume without loss of generality that at
 least one γ_i is positive, and then write

$$\sum_{\{i \mid \gamma_i \leq 0\}} (-\gamma_i)g(x_i) = \sum_{\{i \mid \gamma_i > 0\}} \gamma_i g(x_i) \quad \text{for all } g \in \mathcal{G}. \quad (4.26)$$

13 Now suppose that there exists some $g \in \mathcal{G}$ such that the associated subgraph set
 14 $S_g = \{x \in \mathbb{R}^d \mid g(x) \leq 0\}$ includes only the subset $\{x_i \mid \gamma_i \leq 0\}$. For such a function g ,
 15 the right-hand side of equation (4.26) would be strictly positive while the left-hand side
 16 would be non-positive, which is a contradiction. We conclude that \mathcal{G} fails to shatter
 17 the set $\{x_1, \dots, x_n\}$, as claimed. \square

18 Let us illustrate the use of Proposition 4.5 with some examples:

19 **Example 4.9** (Linear functions in \mathbb{R}^d). For a pair $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, define the function
 20 $f_{a,b}(x) := \langle a, x \rangle + b$, and consider the family $\mathcal{L}^d := \{f_{a,b} \mid (a, b) \in \mathbb{R}^d \times \mathbb{R}\}$ of all such
 21 linear functions. The associated subgraph class $\mathfrak{S}(\mathcal{L}^d)$ corresponds to the collection of
 22 all half-spaces of the form $H_{a,b} := \{x \in \mathbb{R}^d \mid \langle a, x \rangle + b \leq 0\}$. Since the family \mathcal{L}^d
 23 forms a vector space of dimension $d + 1$, we obtain as an immediate consequence of
 24 Proposition 4.5 that $\mathfrak{S}(\mathcal{L}^d)$ has VC dimension at most $d + 1$.

For the special case $d = 1$, let us verify this statement by a more direct calculation.
 In this case, the class $\mathfrak{S}(\mathcal{L}^1)$ corresponds to the collection of all left-sided or right-sided

intervals—that is,

$$\mathbb{S}(\mathcal{L}^1) = \{(-\infty, t] \mid t \in \mathbb{R}\} \cup \{[t, \infty) \mid t \in \mathbb{R}\}.$$

Given any two distinct points $x_1 < x_2$, the collection of all such intervals can pick out all possible subsets. However, given any three points $x_1 < x_2 < x_3$, there is no interval contained in $\mathbb{S}(\mathcal{L}^1)$ that contains x_2 while excluding both x_1 and x_3 . This calculation shows that $\nu(\mathbb{S}(\mathcal{L}^1)) = 2$, which matches the upper bound obtained from Proposition 4.5. More generally, it can be shown that the VC dimension of $\mathbb{S}(\mathcal{L}^d)$ is $d + 1$, so that Proposition 4.5 yields a sharp result in all dimensions. 

Example 4.10 (Spheres in \mathbb{R}^d). Consider the sphere $S_{a,b} := \{x \in \mathbb{R}^d \mid \|x - a\|_2 \leq b\}$, where $(a, b) \in \mathbb{R}^d \times \mathbb{R}_+$ specify its center and radius, respectively, and let $\mathbb{S}_{\text{sphere}}^d$ denote the collection of all such spheres. If we define the function

$$f_{a,b}(x) := \|x\|_2^2 - 2 \sum_{j=1}^d a_j x_j + \|a\|_2^2 - b^2,$$

then we have $S_{a,b} = \{x \in \mathbb{R}^d \mid f_{a,b}(x) \leq 0\}$, so that the sphere $S_{a,b}$ is a sub-graph of the function $f_{a,b}$.

In order to leverage Proposition 4.5, we first define a feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}$ via $\phi(x) := (x_1, \dots, x_d, 1)$, and then consider functions of the form

$$g_c(x) := \langle c, \phi(x) \rangle + \|x\|_2^2, \quad \text{where } c \in \mathbb{R}^{d+1}.$$

The family of functions $\{g_c, c \in \mathbb{R}^{d+1}\}$ is a vector space of dimension $d + 1$, and it contains the function class $\{f_{a,b}, (a, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$. Consequently, by applying Proposition 4.5 to this larger vector space, we conclude that $\nu(\mathbb{S}_{\text{sphere}}^d) \leq d + 2$. This bound is adequate for many purposes, but is not sharp: a more careful analysis shows that the VC dimension of spheres in \mathbb{R}^d is actually $d + 1$. See Exercise 4.9 for an in-depth exploration of the case $d = 2$. 

Appendix A: Proof of Proposition 4.3

We prove inequality (i) in Proposition 4.3 by establishing the following more general inequality:

Lemma 4.2. Let A be a finite set and let \mathcal{U} be a class of subsets of A . Then

$$\text{card}(\mathcal{U}) \leq \text{card}(\{B \subseteq A \mid B \text{ is shattered by } \mathcal{U}\}). \quad (4.27)$$

To see that this lemma implies inequality (i), note that if $B \subseteq A$ is shattered by \mathcal{S} , then we must have $\text{card}(B) \leq \nu(\mathcal{S})$. Consequently, if we let $A = \{x_1, \dots, x_n\}$ and set $\mathcal{U} = \mathcal{S} \cap A$, then Lemma 4.2 implies that

$$\text{card}(\mathcal{S}(x_1^n)) = \text{card}(\mathcal{S} \cap A) \leq \text{card}(\{B \subseteq A \mid |B| \leq \nu(\mathcal{S})\}) \leq \sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i},$$

1 as claimed.

It remains to prove Lemma 4.2. For a given $x \in A$, let us define an operator on sets $U \in \mathcal{U}$ via

$$T_x(U) = \begin{cases} U \setminus \{x\} & \text{if } x \in U \text{ and } U \setminus \{x\} \notin \mathcal{U} \\ U & \text{otherwise.} \end{cases}$$

2 We let $T_x(\mathcal{U})$ be the new class of sets defined by applying T_x to each member of \mathcal{U} —
3 namely, $T_x(\mathcal{U}) := \{T_x(U) \mid U \in \mathcal{U}\}$.

4 We first claim that T_x is a one-to-one mapping between \mathcal{U} and $T_x(\mathcal{U})$, and hence
5 that $\text{card}(T_x(\mathcal{U})) = \text{card}(\mathcal{U})$. To establish this claim, for any pair of sets $U, U' \in \mathcal{U}$ such
6 that $T_x(U) = T_x(U')$, we must prove that $U = U'$. We divide the argument into three
7 separate cases:

8 • *Case 1: $x \notin U$ and $x \notin U'$.* Given $x \notin U$, we have $T_x(U) = U$, and hence $T_x(U') =$
9 U . Moreover, $x \notin U'$ implies that $T_x(U') = U'$. Combining the equalities yields
10 $U = U'$.

11 • *Case 2: $x \notin U$ and $x \in U'$.* In this case, we have $T_x(U) = U = T_x(U')$, so that
12 $x \in U'$ but $x \notin T_x(U')$. But this condition implies that $T_x(U') = U' \setminus \{x\} \notin \mathcal{U}$,
13 which contradicts the fact that $T_x(U') = U \in \mathcal{U}$. By symmetry, the case $x \in U$
14 and $x \notin U'$ is identical.

15 • *Case 3: $x \in U \cap U'$.* If both of $U \setminus \{x\}$ and $U' \setminus \{x\}$ belong to \mathcal{U} , then $T_x(U) = U$ and
16 $T_x(U') = U'$, from which $U = U'$ follows. If neither of $U \setminus \{x\}$ nor $U' \setminus \{x\}$ belong
17 to \mathcal{U} , then we can conclude that $U \setminus \{x\} = U' \setminus \{x\}$, and hence $U = U'$. Finally,
18 if $U \setminus \{x\} \notin \mathcal{U}$ but $U' \setminus \{x\} \in \mathcal{U}$, then $T_x(U) = U \setminus \{x\} \notin \mathcal{U}$ but $T_x(U') = U' \in \mathcal{U}$,
19 which is a contradiction.

20 We next claim that if $T_x(\mathcal{U})$ shatters a set B , then so does \mathcal{U} . If $x \notin B$, then both \mathcal{U}
21 and $T_x(\mathcal{U})$ pick out the same set of subsets of B . Otherwise, suppose that $x \in B$. Since
22 $T_x(\mathcal{U})$ shatters B , for any subset $B' \subseteq B \setminus \{x\}$, there is a subset $T \in T_x(\mathcal{U})$ such that
23 $T \cap B = B' \cup \{x\}$. Since $T = T_x(U)$ for some subset $U \in \mathcal{U}$ and $x \in T$, we conclude
24 that both U and $U \setminus \{x\}$ must belong to \mathcal{U} , so that \mathcal{U} also shatters B .

25 Using the fact that T_x preserves cardinalities and does not increase shattering
26 numbers, we can now conclude the proof of the lemma. Define the weight function

$\omega(\mathcal{U}) = \sum_{U \in \mathcal{U}} \text{card}(U)$, and note that applying a transformation T_x can only reduce this weight function: i.e., $\omega(T_x(\mathcal{U})) \leq \omega(\mathcal{U})$. Consequently, by applying the transformations $\{T_x\}$ to \mathcal{U} repeatedly, we can obtain a new class of sets \mathcal{U}' such that $\text{card}(\mathcal{U}) = \text{card}(\mathcal{U}')$ and the weight $\omega(\mathcal{U}')$ is minimal. Then for any $U \in \mathcal{U}'$ and any $x \in U$, we have $U \setminus \{x\} \in \mathcal{U}'$. (Otherwise, we would have $\omega(T_x(\mathcal{U}')) < \omega(\mathcal{U}')$, contradicting minimality.) Therefore, the set class \mathcal{U}' shatters any one of its elements. Since \mathcal{U} shatters at least as many subsets as \mathcal{U}' , and $\text{card}(\mathcal{U}') = \text{card}(\mathcal{U})$ by construction, the claim (4.27) follows.

■ 4.4 Bibliographic details and background

First, a technical remark regarding measurability: in general, the quantity $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ need not be measurable, since the function class \mathcal{F} may contain an uncountable number of elements. If the function class is separable, then we may simply take the supremum over the countable dense basis. In general, there are various ways of dealing with this issue, including the use of outer probability (c.f. van der Vaart and Wellner [vdVW96]). Here we instead adopt the following convention. For any finite class of functions \mathcal{G} contained within \mathcal{F} , the variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}}$ is well-defined, so that it is sensible to define

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup \{ \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}} \mid \mathcal{G} \subset \mathcal{F}, \mathcal{G} \text{ has finite cardinality} \}.$$

By using this definition, we can always think instead about suprema over finite sets.

Theorem 4.1 was originally proved by Glivenko [Gli33] for the continuous case, and by Cantelli [Can33] in the general setting. The non-asymptotic form of the Glivenko-Cantelli theorem given in Corollary 4.1 can be sharpened substantially. For instance, Dvoretzky, Kiefer and Wolfowitz [DKW56] prove that there is a constant C independent of F and n such that

$$\mathbb{P}[\|\widehat{F}_n - F\|_{\infty} \geq \delta] \leq C e^{-2n\delta^2} \quad \text{for all } \delta \geq 0. \quad (4.28)$$

Massart [Mas90] establishes the sharpest possible result, including control of the leading constant.

The Rademacher complexity, and its relative the Gaussian complexity, have a lengthy history in the study of Banach spaces using probabilistic methods; for instance, see the books [MS86, Pis89, LT91]. In Chapter 5, we develop further connections between these two forms of complexity, and the related notion of metric entropy. Rademacher and Gaussian complexities have also been studied in the specific context of uniform laws of large numbers and empirical risk minimization (e.g., [BM02, BBM05, Kol01, Kol06, KP00, vdVW96]).

The proof of Proposition 4.3 is adapted from Ledoux and Talagrand [Led01], who credit the approach to Frankl [Fra83]. Exercise 5.4 is adapted from Problem 2.6.3 from

1 van der Vaart and Wellner [vdVW96]. The proof of Proposition 4.5 is adapted from
 2 Pollard [Pol84], who credits it to Steele [Ste78] and Dudley [Dud78].

3 ■ 4.5 Exercises

4 **Exercise 4.1.** Recall that the functional γ is *continuous in the sup-norm at F* if for
 5 all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|G - F\|_\infty \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \epsilon$.

6 (a) Given n i.i.d. samples with law specified by F , let \widehat{F}_n be the empirical CDF. Show
 7 that if γ is continuous in the sup-norm at F , then $\gamma(\widehat{F}_n) \xrightarrow{\text{prob.}} \gamma(F)$.

8 (b) Which of the following functionals are continuous with respect to the sup-norm?
 9 Prove or disprove.

10 (i) The mean functional $F \mapsto \int x dF(x)$.

11 (ii) The Cramér-von Mises functional $F \mapsto \int [F(x) - F_0(x)]^2 dF_0(x)$.

12 (iii) The quantile functional $Q_p(F) = \inf\{t \in \mathbb{R} \mid F(t) \geq p\}$.

Exercise 4.2. Recall from Example 4.5 the class \mathcal{S} of all subsets S of $[0, 1]$ for which
 S has a finite number of elements. Prove that the Rademacher complexity satisfies the
 lower bound

$$\mathcal{R}_n(\mathcal{S}) = \mathbb{E}_{X,\varepsilon} \left[\sup_{S \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbb{1}_S[X_i] \right| \right] \geq \frac{1}{2}. \quad (4.29)$$

13 Discuss the connection to Theorem 4.2.

14 **Exercise 4.3.** In this exercise, we work through the proof of Proposition 4.2.

(a) Recall the re-centered function class $\overline{\mathcal{F}} = \{f - \mathbb{E}[f] \mid f \in \mathcal{F}\}$. Show that

$$\mathbb{E}_{X,\varepsilon}[\|R_n\|_{\overline{\mathcal{F}}}] \geq \mathbb{E}_{X,\varepsilon}[\|R_n\|_{\mathcal{F}}] - \frac{\sup_{f \in \mathcal{F}} |\mathbb{E}[f]|}{\sqrt{n}}.$$

15 (b) Use concentration results to complete the proof of Proposition 4.2.

Exercise 4.4. Consider the function class

$$\mathcal{F} = \{x \mapsto \text{sign}(\langle \theta, x \rangle) \mid \theta \in \mathbb{R}^d, \|\theta\|_2 = 1\},$$

corresponding to the $\{-1, +1\}$ -valued classification rules defined by linear functions in
 \mathbb{R}^d . Supposing that $d \geq n$, let $x_1^n = \{x_1, \dots, x_n\}$ be a collection of vectors in \mathbb{R}^d that

are linearly independent. Show that the empirical Rademacher complexity satisfies

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] = 1.$$

Discuss the consequences for empirical risk minimization over the class \mathcal{F} .

Exercise 4.5. Prove the following properties of the Rademacher complexity.

(a) $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\text{conv}(\mathcal{F}))$.

(b) Show that $\mathcal{R}_n(\mathcal{F} + \mathcal{G}) \leq \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$. Give an example to demonstrate that this bound cannot be improved in general.

(c) Given a fixed and uniformly bounded function g , show that

$$\mathcal{R}_n(\mathcal{F} + g) \leq \mathcal{R}_n(\mathcal{F}) + \frac{\|g\|_\infty}{\sqrt{n}}. \quad (4.30)$$

Exercise 4.6. Let \mathbb{S} and \mathbb{T} be two classes of sets with finite VC dimensions. Which of the following set classes have finite VC dimension? Prove or disprove.

(a) The set class $\mathbb{S}^c := \{S^c \mid S \in \mathbb{S}\}$, where S^c denotes the complement of the set S .

(b) The set class $\mathbb{S} \cap \mathbb{T} := \{S \cap T \mid S \in \mathbb{S}, T \in \mathbb{T}\}$.

(c) The set class $\mathbb{S} \sqcup \mathbb{T} := \{S \cup T \mid S \in \mathbb{S}, T \in \mathbb{T}\}$.

Exercise 4.7. Prove Lemma 4.1.

Exercise 4.8. Consider the class of left-sided half-intervals in \mathbb{R}^d :

$$\mathbb{S}_{\text{left}}^d := \{(-\infty, t_1] \times (-\infty, t_2] \times \cdots \times (-\infty, t_d] \mid (t_1, \dots, t_d) \in \mathbb{R}^d\}.$$

Show that for any collection of n points, we have $\text{card}(\mathbb{S}_{\text{left}}^d(x_1^n)) \leq (n+1)^d$ and $\nu(\mathbb{S}_{\text{left}}^d) = d$.

Exercise 4.9. Consider the class of all spheres in \mathbb{R}^2 : —that is

$$\mathbb{S}_{\text{sphere}}^2 := \{S_{a,b}, (a,b) \in \mathbb{R}^2 \times \mathbb{R}_+\}, \quad (4.31)$$

where $S_{a,b} := \{x \in \mathbb{R}^2 \mid \|x - a\|_2 \leq b\}$ is the sphere of radius $b \geq 0$ centered at $a = (a_1, a_2)$.

(a) Show that $\mathbb{S}_{\text{sphere}}^2$ can shatter any subset of three points that are not collinear.

- 1 (b) Show that for any subset of four points, the balls can discriminate at most 15 out
2 of 16 of the possible subsets, and conclude that $\nu(\mathbb{S}_{\text{sphere}}^2) = 3$.
- 3 **Exercise 4.10.** Show that the class \mathbb{C}_{cc}^d of all closed and convex sets in \mathbb{R}^d does *not*
4 have finite VC dimension. (*Hint:* Consider a set of n points on the boundary of the
5 unit ball.)
- 6 **Exercise 4.11.** (a) Prove inequality (ii) in Proposition 4.3.
- 7 (b) Prove the sharper upper bound $\text{card}(\mathbb{S}(x_1^n)) \leq \left(\frac{en}{\nu}\right)^\nu$. (*Hint:* You might find the
8 result of Exercise 2.9 useful.)