4

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 entry point 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.

4.1 Motivation

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

4.1.1 Uniform convergence of cumulative distribution functions

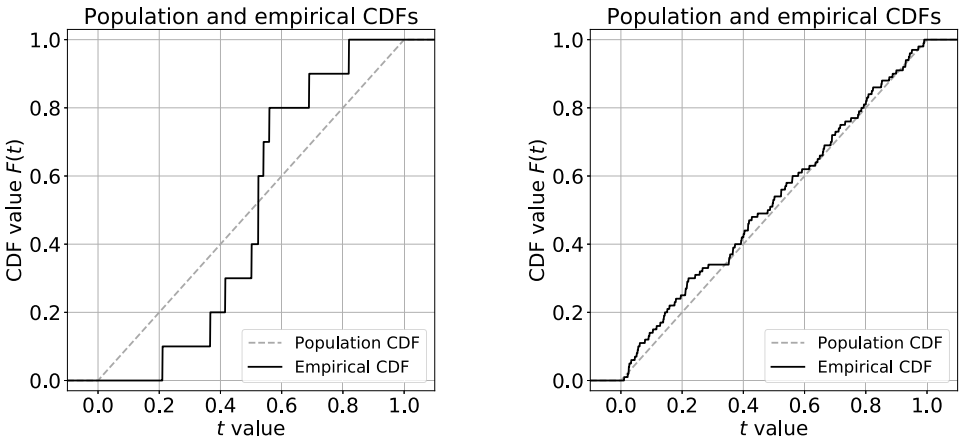
The law of any scalar random variable X can be fully specified by its cumulative distribution function (CDF), whose value at any point $t \in \mathbb{R}$ is given by $F(t) := \mathbb{P}[X \leq t]$. Now suppose that we are given a collection $\{X_i\}_{i=1}^n$ of n i.i.d. samples, each drawn according to the law specified by F. A natural estimate of F is the *empirical CDF* given by

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

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

Figure 4.1 provides some illustrations of empirical CDFs for the uniform distribution on the interval $[0,1]$ for two different sample sizes. Note that $\widehat{F}_n$ is a random function, with the value $\widehat{F}_n(t)$ corresponding to the fraction of samples that lie in the interval $(-\infty, t]$. As the sample size n grows, we see that $\widehat{F}_n$ approaches F—compare the plot for $n = 10$ in Figure 4.1(a) to that for $n = 100$ in Figure 4.1(b). It is easy to see that $\widehat{F}_n$ converges to F in

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press



(a) (b)

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.

a pointwise sense. Indeed, for any fixed $t \in \mathbb{R}$, the random variable $\widehat{F}_n(t)$ has mean $F(t)$, and moments of all orders, so that the strong law of large numbers implies that $\widehat{F}_n(t) \xrightarrow{a.s.} F(t)$. A natural goal is to strengthen this pointwise convergence to a form of uniform convergence.

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

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

$$\gamma_g(F) := \int g(x)\, dF(x). \tag{4.2}$$

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

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

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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

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

and corresponds to estimating the αth quantile of the distribution by the αth 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, nonlinear 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 case 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—that is, when does $\gamma(\widehat{F}_n)$ 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 CDFs F and G, let us measure the distance between them using the sup-norm

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

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 the random variable $\|\widehat{F}_n - F\|_\infty$ converges to zero. A classical result, known as the Glivenko–Cantelli theorem, addresses the latter question:

Theorem 4.4 (Glivenko–Cantelli) *For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that*

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

We provide a proof of this claim as a corollary of a more general result to follow (see Theorem 4.10). For statistical applications, an important consequence of Theorem 4.4 is

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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 this connection.

4.1.2 Uniform laws for more general function classes

We now turn to more general consideration of uniform laws of large numbers. Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable

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

which measures the absolute deviation between the sample average $\frac{1}{n}\sum_{i=1}^n f(X_i)$ and the population average $\mathbb{E}[f(X)]$, uniformly over the class $\mathscr{F}$. Note that there can be measurability concerns associated with the definition (4.7); see the bibliographic section for discussion of different ways in which to resolve them.

Definition 4.5 We say that $\mathscr{F}$ is a *Glivenko–Cantelli* class for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ converges to zero in probability as $n \to \infty$.

This notion can also be defined in a stronger sense, requiring almost sure convergence of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$, in which case we say that $\mathscr{F}$ satisfies a *strong Glivenko–Cantelli law.* The classical result on the empirical CDF (Theorem 4.4) can be reformulated as a particular case of this notion:

Example 4.6 (Empirical CDFs and indicator functions) Consider the function class

$$\mathscr{F} = \{\mathbb{I}_{(-\infty,t]}(\cdot) \mid t \in \mathbb{R}\}, \tag{4.8}$$

where $\mathbb{I}_{(-\infty,t]}$ is the $\{0,1\}$-valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have the equality $\mathbb{E}[\mathbb{I}_{(-\infty,t]}(X)] = \mathbb{P}[X \leq t] = F(t)$, so that the classical Glivenko–Cantelli theorem is equivalent 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 4.7 (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 $\mathscr{F}_{\mathcal{S}} = \{\mathbb{I}_S(\cdot) \mid S \in \mathcal{S}\}$ 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 discrete set $\{X_1, \ldots, X_n\}$ belongs to $\mathcal{S}$, and moreover, by definition of the empirical distribution, we have $\mathbb{P}_n[X_1^n] = 1$. Putting together

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

the pieces, we conclude that

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

so that the function class $\mathscr{F}_{\mathcal{S}}$ 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 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$—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 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 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 $\{X_i\}_{i=1}^n$, each sample lying in some space $\mathcal{X}$. Suppose that the samples are drawn i.i.d. according to a distribution $\mathbb{P}_{\theta^*}$, 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 $\Omega = \mathscr{G}$, in which case the problem is of the nonparametric variety.

In either case, a standard decision-theoretic approach to estimating θ^* is based on minimizing a cost function of the form $\theta \mapsto \mathcal{L}_\theta(X)$, which measures the "fit" between a parameter $\theta \in \Omega$ and the sample $X \in \mathcal{X}$. Given the collection of n samples $\{X_i\}_{i=1}^n$, 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 $\mathbb{P}_{\theta^*}$. This empirical risk should be contrasted with the *population risk*,

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

where the expectation $\mathbb{E}_{\theta^*}$ 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

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

Let us consider some examples to illustrate.

Example 4.8 (Maximum likelihood) Consider a parameterized family of distributions—say $\{\mathbb{P}_\theta, \theta \in \Omega\}$—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 an unknown distribution $\mathbb{P}_{\theta^*}$, and we would like to estimate the unknown parameter θ^*. In order to do so, we consider the cost function

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

The term $p_{\theta^*}(x)$, which we have included for later theoretical convenience, has no effect on

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

the minimization over θ. Indeed, the maximum likelihood estimate is obtained by minimizing the empirical risk defined by this cost function—that is

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

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 *Kullback–Leibler divergence* between p_{θ^*} and p_θ. In the special case that $\theta^* \in \Omega_0$, the excess risk is simply the Kullback–Leibler divergence between the true density p_{θ^*} and the fitted model $p_{\widehat{\theta}}$. See Exercise 4.3 for some concrete examples. ♣

Example 4.9 (Binary classification) Suppose that we observe n pairs of samples, each of the form $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, +1\}$, 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 $\mathbb{P}_{Y|X}$. 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 $f : \mathbb{R}^d \to \{-1, +1\}$ that minimizes the probability of misclassification $\mathbb{P}[f(X) \neq Y]$, for an independently drawn pair (X, Y). Note that this probability of error corresponds to the population risk for the cost function

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

A function that minimizes this probability of error is known as a *Bayes classifier* f^*; in the special case of equally probable classes—that is, when $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = \frac{1}{2}$—a Bayes classifier is given by

$$f^*(x) = \begin{cases} +1 & \text{if } \psi(x) \geq 1, \\ -1 & \text{otherwise.} \end{cases}$$

Since the likelihood ratio ψ (and hence f^*) is unknown, a natural approach to approximating the Bayes rule is based on choosing $\widehat{f}$ to minimize the empirical risk

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

corresponding to the fraction of training samples that are misclassified. 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 assume[1] that there exists some $\theta_0 \in \Omega_0$ such that $R(\theta_0, \theta^*) =$

[1] If 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.

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

$\inf_{\theta \in \Omega_0} R(\theta, \theta^*)$. With this notation, the excess risk can be decomposed as

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

Note that T_2 is non-positive, since $\widehat{\theta}$ minimizes the empirical risk over Ω_0.

The third term T_3 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 empirical risk, we have

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

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 cost function is bounded.

Finally, returning to the first term, it can be written in a similar way, namely as the difference

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

This quantity is more challenging to control, because the parameter $\widehat{\theta}$—in contrast to the deterministic quantity θ_0—is now random, and moreover depends on the samples $\{X_i\}_{i=1}^n$, 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 cost function class $\mathfrak{L}(\Omega_0) := \{x \mapsto \mathcal{L}_\theta(x), \theta \in \Omega_0\}$. With this notation, we have

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

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

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 $\mathscr{F}$. For any fixed collection $x_1^n := (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathscr{F}(x_1^n) := \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathscr{F}\}. \tag{4.11}$$

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

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

given by

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

Note that this definition coincides with our earlier definition of the Rademacher complexity of a set (see Example 2.25).

Given a collection $X_1^n := \{X_i\}_{i=1}^n$ of random samples, then the empirical Rademacher complexity $\mathcal{R}(\mathscr{F}(X_1^n)/n)$ is a random variable. Taking its expectation yields the *Rademacher complexity of the function class* $\mathscr{F}$—namely, the deterministic quantity

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

Note that the Rademacher complexity is the average of the maximum correlation between the vector $(f(X_1), \ldots, f(X_n))$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$, where the maximum is taken over all functions $f \in \mathscr{F}$. 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, for any bounded function class $\mathscr{F}$, the condition $\mathcal{R}_n(\mathscr{F}) = o(1)$ implies the Glivenko–Cantelli property. More precisely, we prove a non-asymptotic statement, in terms of a tail bound for the probability that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ deviates substantially above a multiple of the Rademacher complexity. It applies to a function class $\mathscr{F}$ that is b-uniformly bounded, meaning that $\|f\|_\infty \leq b$ for all $f \in \mathscr{F}$.

Theorem 4.10 *For any b-uniformly bounded class of functions $\mathscr{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have*

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

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

In order for Theorem 4.10 to be useful, we need to obtain upper bounds on the Rademacher complexity. There are a variety of methods for doing so, ranging from direct calculations to alternative complexity measures. In Section 4.3, we develop some techniques for upper bounding the Rademacher complexity for indicator functions of half-intervals, as required for the classical Glivenko–Cantelli theorem (see Example 4.6); we also discuss the notion of Vapnik–Chervonenkis dimension, which can be used to upper bound the Rademacher complexity for other function classes. In Chapter 5, we introduce more advanced

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

techniques based on metric entropy and chaining for controlling Rademacher complexity and related sub-Gaussian processes. In the meantime, let us turn to the proof of Theorem 4.10.

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

Concentration around mean: We first claim that, when $\mathscr{F}$ is uniformly bounded, then the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ is sharply concentrated around its mean. In order to simplify notation, it is convenient to define the recentered functions $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, and to write $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(X_i) \right|$. Thinking of the samples as fixed for the moment, consider the function

$$G(x_1, \ldots, x_n) := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|.$$

We claim that G satisfies the Lipschitz property required to apply the bounded differences method (recall Corollary 2.21). Since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first coordinate 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 to bound the difference $|G(x) - G(y)|$. For any function $\bar{f} = f - \mathbb{E}[f]$, we have

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

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 $f \in \mathscr{F}$ 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 2.21), we have

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

valid for all $t \geq 0$.

Upper bound on mean: It remains to show that $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}]$ is upper bounded by $2\mathcal{R}_n(\mathscr{F})$, and we do so using a classical symmetrization argument. Letting $(Y_1, \ldots, Y_n)$ be a second

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

i.i.d. sequence, independent of $(X_1, \ldots, X_n)$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}] &= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}\right|\right] \\ &= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}_Y[\frac{1}{n}\sum_{i=1}^n\{f(X_i) - f(Y_i)\}]\right|\right] \\ &\overset{(i)}{\leq} \mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\{f(X_i) - f(Y_i)\}\right|\right], \end{aligned} \tag{4.17}$$

where the upper bound (i) follows from the calculation of Exercise 4.4.

Now let $(\varepsilon_1, \ldots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y. Given our independence assumptions, for any function $f \in \mathscr{F}$, the random vector with components $\varepsilon_i(f(X_i) - f(Y_i))$ has the same joint distribution as the random vector with components $f(X_i) - f(Y_i)$, whence

$$\begin{aligned}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\{f(X_i) - f(Y_i)\}\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(X_i) - f(Y_i))\right|\right] \\ &\leq 2\,\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] = 2\,\mathcal{R}_n(\mathscr{F}). \end{aligned} \tag{4.18}$$

Combining the upper bound (4.18) with the tail bound (4.16) yields the claim. □

4.2.1 Necessary conditions with Rademacher complexity

The proof of Theorem 4.10 illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ to its symmetrized version

$$\|\mathbb{S}_n\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|. \tag{4.19}$$

Note that the expectation of $\|\mathbb{S}_n\|_{\mathscr{F}}$ corresponds to the Rademacher complexity, which plays a central role in Theorem 4.10. It is natural to wonder whether much was lost in moving from the variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ to its symmetrized version. The following "sandwich" result relates these quantities.

Proposition 4.11 *For any convex non-decreasing function $\Phi\colon \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}_{X,\varepsilon}[\Phi(\tfrac{1}{2}\|\mathbb{S}_n\|_{\mathscr{F}})] \overset{(a)}{\leq} \mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}})] \overset{(b)}{\leq} \mathbb{E}_{X,\varepsilon}[\Phi(2\|\mathbb{S}_n\|_{\mathscr{F}})], \tag{4.20}$$

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

When applied with the convex non-decreasing function $\Phi(t) = t$, Proposition 4.11 yields the inequalities

$$\tfrac{1}{2}\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\bar{\mathscr{F}}} \leq \mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}] \leq 2\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\mathscr{F}}, \tag{4.21}$$

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

with the only differences being the constant pre-factors, and the use of $\mathscr{F}$ in the upper bound, and the recentered class $\overline{\mathscr{F}}$ 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 Beginning with bound (b), we have

$$\begin{aligned}
\mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}})] &= \mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_Y[f(Y_i)]\right|\right)\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right)\right] \\
&\overset{\text{(ii)}}{=} \underbrace{\mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{f(X_i) - f(Y_i)\}\right|\right)\right]}_{:=T_1},
\end{aligned}$$

where inequality (i) follows from Exercise 4.4, using the convexity and non-decreasing properties of Φ, and equality (ii) follows since the random vector with components $\varepsilon_i(f(X_i) - f(Y_i))$ has the same joint distribution as the random vector with components $f(X_i) - f(Y_i)$. By the triangle inequality, we have

$$\begin{aligned}
T_1 &\leq \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(iii)}}{\leq} \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{Y,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(iv)}}{=} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right],
\end{aligned}$$

where step (iii) follows from Jensen's inequality and the convexity of Φ, and step (iv) follows since X and Y are i.i.d. samples.

Turning to the bound (a), we have

$$\begin{aligned}
\mathbb{E}_{X,\varepsilon}[\Phi(\tfrac{1}{2}\|\mathbb{S}_n\|_{\mathscr{F}})] &= \mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}\right|\right)\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{f(X_i) - f(Y_i)\}\right|\right)\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(Y_i)\}\right|\right)\right],
\end{aligned}$$

where inequality (i) follows from Jensen's inequality and the convexity of Φ; and equality (ii) follows since for each $i = 1, 2, \ldots, n$ and $f \in \mathscr{F}$, the variables $\varepsilon_i\{f(X_i) - f(Y_i)\}$ and $f(X_i) - f(Y_i)$ have the same distribution.

Now focusing on the quantity $T_2 := \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\{f(X_i) - f(Y_i)\}\right|$, we add and subtract a term of the form $\mathbb{E}[f]$, and then apply the triangle inequality, thereby obtaining the upper

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

bound

$$T_2 \leq \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-\mathbb{E}[f]\}\right| + \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(Y_i)-\mathbb{E}[f]\}\right|.$$

Since Φ is convex and non-decreasing, we are guaranteed that

$$\Phi(T_2) \leq \frac{1}{2}\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-\mathbb{E}[f]\}\right|\right) + \frac{1}{2}\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(Y_i)-\mathbb{E}[f]\}\right|\right).$$

The claim follows by taking expectations and using the fact that X and Y are identically distributed. □

A consequence of Proposition 4.11 is that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ can be lower bounded by a multiple of Rademacher complexity, and some fluctuation terms. This fact can be used to prove the following:

Proposition 4.12 *For any b-uniformly bounded function class $\mathscr{F}$, any integer $n \geq 1$ and any scalar $\delta \geq 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}[f]|}{2\sqrt{n}} - \delta \tag{4.22}$$

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

We leave the proof of this result for the reader (see Exercise 4.5). As a consequence, if the Rademacher complexity $\mathcal{R}_n(\mathscr{F})$ remains bounded away from zero, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ cannot converge to zero in probability. We have thus shown that, for a uniformly bounded function class $\mathscr{F}$, 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 4.10 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, as well as associated Vapnik–Chervonenkis classes.

4.3.1 Classes with polynomial discrimination

For a given collection of points $x_1^n = (x_1, \ldots, x_n)$, the "size" of the set $\mathscr{F}(x_1^n)$ provides a sample-dependent measure of the complexity of $\mathscr{F}$. In the simplest case, the set $\mathscr{F}(x_1^n)$ con-

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

tains only a finite number of vectors for all sample sizes, so that its "size" can be measured via its cardinality. For instance, if $\mathscr{F}$ consists of a family of decision rules taking binary values (as in Example 4.9), then $\mathscr{F}(x_1^n)$ 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 4.13 (Polynomial discrimination) A class $\mathscr{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$ if, for each positive integer n and collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathcal{X}$, the set $\mathscr{F}(x_1^n)$ has cardinality upper bounded as

$$\operatorname{card}(\mathscr{F}(x_1^n)) \leq (n+1)^{\nu}. \tag{4.23}$$

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

Lemma 4.14 *Suppose that $\mathscr{F}$ has polynomial discrimination of order ν. Then for all positive integers n and any collection of points $x_1^n = (x_1, \ldots, x_n)$,*

$$\underbrace{\mathbb{E}_{\varepsilon}\left[\sup_{f \in \mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)\right|\right]}_{\mathcal{R}(\mathscr{F}(x_1^n)/n))} \leq 4D(x_1^n)\sqrt{\frac{\nu \log(n+1)}{n}},$$

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

We leave the proof of this claim for the reader (see Exercise 4.9).

Although Lemma 4.14 is stated as an upper bound on the empirical Rademacher complexity, it yields as a corollary an upper bound on the Rademacher complexity $\mathcal{R}_n(\mathscr{F}) = \mathbb{E}_X[\mathcal{R}(\mathscr{F}(X_1^n)/n)]$, one which involves the expected ℓ_2-width $\mathbb{E}_{X_1^n}[D(X)]$. An especially simple case is when the function class is b uniformly bounded, so that $D(x_1^n) \leq b$ for all samples. In this case, Lemma 4.14 implies that

$$\mathcal{R}_n(\mathscr{F}) \leq 2b\sqrt{\frac{\nu \log(n+1)}{n}} \qquad \text{for all } n \geq 1. \tag{4.24}$$

Combined with Theorem 4.10, 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 4.6, the classical Glivenko–Cantelli law is based on indicator functions of the

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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 $\nu = 1$. Consequently, Theorem 4.10 combined with Lemma 4.14 yields a quantitative version of Theorem 4.4 as a corollary.

Corollary 4.15 (Classical Glivenko–Cantelli) *Let $F(t) = \mathbb{P}[X \leq t]$ 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*

$$\mathbb{P}\left[\|\widehat{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \leq e^{-\frac{n\delta^2}{2}} \qquad \text{for all } \delta \geq 0, \tag{4.25}$$

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

Proof For a given sample $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^n$, consider the set $\mathscr{F}(x_1^n)$, where $\mathscr{F}$ is the set of all {0-1}-valued indicator functions of the half-intervals $(-\infty, t]$ for $t \in \mathbb{R}$. If we order the samples as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, 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 $\mathbb{I}_{(-\infty,t]}$ takes the value one for all $x_{(i)} \leq t$, and the value zero for all other samples. Thus, we have shown that, for any given sample x_1^n, we have $\operatorname{card}(\mathscr{F}(x_1^n)) \leq n+1$. Applying Lemma 4.14, we obtain

$$\mathbb{E}_\varepsilon\left[\sup_{f \in \mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \leq 4\sqrt{\frac{\log(n+1)}{n}},$$

and taking averages over the data X_i yields the upper bound $\mathcal{R}_n(\mathscr{F}) \leq 4\sqrt{\frac{\log(n+1)}{n}}$. The claim (4.25) then follows from Theorem 4.10. □

Although the exponential tail bound (4.25) 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 in a less laborious manner. The theory of Vapnik–Chervonenkis (VC) dimension provides one such class of techniques. Accordingly, we now turn to defining the notions of shattering and VC dimension.

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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

Definition 4.16 (Shattering and VC dimension) Given a class $\mathscr{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \ldots, x_n)$ is *shattered* by $\mathscr{F}$ if $\text{card}(\mathscr{F}(x_1^n)) = 2^n$. The *VC dimension* $\nu(\mathscr{F})$ is the largest integer n for which there is *some* collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathscr{F}$.

When the quantity $\nu(\mathscr{F})$ is finite, then the function class $\mathscr{F}$ is said to be a *VC class*. We will frequently consider function classes $\mathscr{F}$ that consist of indicator functions $\mathbb{I}_S[\cdot]$, 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 $\mathscr{F}(x_1^n)$ and the VC dimension of $\mathscr{F}$, respectively. For a given set class $\mathcal{S}$, the shatter coefficient of order n is given by $\max_{x_1^n} \text{card}(\mathcal{S}(x_1^n))$.

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

Example 4.17 (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 4.15 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 4.15, we showed more specifically that, for any collection $x_1^n = \{x_1, \ldots, x_n\}$, 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 $\mathcal{S}_{\text{two}}$ 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 shatter coefficients of $\mathcal{S}_{\text{two}}$. Note that any collection of n distinct points $x_1 < x_2 < \cdots < x_{n-1} < x_n$ 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 of order n is

$$\text{card}(\mathcal{S}_{\text{two}}(x_1^n)) \leq (n + 1)^2,$$

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 also to have 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) in papers by Vapnik and Chervonenkis, Sauer and Shelah.

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 $n > \nu(\mathcal{S})$, then it must be the case that

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

$\text{card}(\mathcal{S}(x_1^n)) < 2^n$ *for all* collections x_1^n of n samples. However, at least in principle, there could exist some subset with

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

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

Proposition 4.18 (Vapnik–Chervonenkis, Sauer and Shelah) *Consider a set class $\mathcal{S}$ with $\nu(\mathcal{S}) < \infty$. Then for any collection of points $P = (x_1, \ldots, x_n)$ with $n \geq \nu(\mathcal{S})$, we have*

$$\text{card}(\mathcal{S}(P)) \overset{\text{(i)}}{\leq} \sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \overset{\text{(ii)}}{\leq} (n+1)^{\nu(\mathcal{S})}. \tag{4.26}$$

Given inequality (i), inequality (ii) can be established by elementary combinatorial arguments, so we leave it to the reader (in particular, see part (a) of Exercise 4.11). Part (b) of the same exercise establishes a sharper upper bound.

Proof Given a subset of points Q and a set class $\mathcal{T}$, we let $\nu(\mathcal{T}; Q)$ denote the VC dimension of $\mathcal{T}$ when considering only whether or not subsets of Q can be shattered. Note that $\nu(\mathcal{T}) \leq k$ implies that $\nu(\mathcal{T}; Q) \leq k$ for all point sets Q. For positive integers (n, k), define the functions

$$\Phi_k(n) := \sup_{\substack{\text{point sets } Q \\ \text{card}(Q) \leq n}} \sup_{\substack{\text{set classes } \mathcal{T} \\ \nu(\mathcal{T}; Q) \leq k}} \text{card}(\mathcal{T}(Q)) \quad \text{and} \quad \Psi_k(n) := \sum_{i=0}^{k} \binom{n}{i}.$$

Here we agree that $\binom{n}{i} = 0$ whenever $i > n$. In terms of this notation, we claim that it suffices to prove that

$$\Phi_k(n) \leq \Psi_k(n). \tag{4.27}$$

Indeed, suppose there were some set class $\mathcal{S}$ with $\nu(\mathcal{S}) = k$ and collection $P = \{x_1, \ldots, x_n\}$ of n distinct points for which $\text{card}(\mathcal{S}(P)) > \Psi_k(n)$. By the definition $\Phi_k(n)$, we would then have

$$\Phi_k(n) \overset{\text{(i)}}{\geq} \sup_{\substack{\text{set classes } \mathcal{T} \\ \nu(\mathcal{T}; P) \leq k}} \text{card}(\mathcal{T}(P)) \overset{\text{(ii)}}{\geq} \text{card}(\mathcal{S}(P)) > \Psi_k(n), \tag{4.28}$$

which contradicts the claim (4.27). Here inequality (i) follows because P is feasible for the supremum over Q that defines $\Phi_k(n)$; and inequality (ii) follows because $\nu(\mathcal{S}) = k$ implies that $\nu(\mathcal{S}; P) \leq k$.

We now prove the claim (4.27) by induction on the sum $n + k$ of the pairs (n, k).

Base case: To start, we claim that inequality (4.27) holds for all pairs with $n + k = 2$.

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

The claim is trivial if either $n = 0$ or $k = 0$. Otherwise, for $(n, k) = (1, 1)$, both sides of inequality (4.27) are equal to 2.

Induction step: Now assume that, for some integer $\ell > 2$, the inequality (4.27) holds for all pairs with $n + k < \ell$. We claim that it then holds for all pairs with $n + k = \ell$. Fix an arbitrary pair (n, k) such that $n + k = \ell$, a point set $P = \{x_1, \ldots, x_n\}$ and a set class $\mathcal{S}$ such that $\nu(\mathcal{S}; P) = k$. Define the point set $P' = P \setminus \{x_1\}$, and let $\mathcal{S}_0 \subseteq \mathcal{S}$ be the smallest collection of subsets that labels the point set P' in the maximal number of different ways. Let $\mathcal{S}_1$ be the smallest collection of subsets inside $\mathcal{S} \setminus \mathcal{S}_0$ that produce binary labelings of the point set P that are not in $\mathcal{S}_0(P)$. (The choices of $\mathcal{S}_0$ and $\mathcal{S}_1$ need not be unique.)

As a concrete example, given a set class $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ and a point set $P = \{x_1, x_2, x_3\}$, suppose that the sets generated the binary labelings

$$s_1 \leftrightarrow (0, 1, 1), \quad s_2 \leftrightarrow (1, 1, 1), \quad s_3 \leftrightarrow (0, 1, 0), \quad s_4 \leftrightarrow (0, 1, 1).$$

In this particular case, we have $\mathcal{S}(P) = \{(0, 1, 1), (1, 1, 1), (0, 1, 0)\}$, and one valid choice of the pair $(\mathcal{S}_0, \mathcal{S}_1)$ would be $\mathcal{S}_0 = \{s_1, s_3\}$ and $\mathcal{S}_1 = \{s_2\}$, generating the labelings $\mathcal{S}_0(P) = \{(0, 1, 1), (0, 1, 0)\}$ and $\mathcal{S}_1(P) = \{(1, 1, 1)\}$.

Using this decomposition, we claim that

$$\operatorname{card}(\mathcal{S}(P)) = \operatorname{card}(\mathcal{S}_0(P')) + \operatorname{card}(\mathcal{S}_1(P')).$$

Indeed, any binary labeling in $\mathcal{S}(P)$ is either mapped to a member of $\mathcal{S}_0(P')$, or in the case that its labeling on P' corresponds to a duplicate, it can be uniquely identified with a member of $\mathcal{S}_1(P')$. This can be verified in the special case described above.

Now since P' is a subset of P and $\mathcal{S}_0$ is a subset of $\mathcal{S}$, we have

$$\nu(\mathcal{S}_0; P') \leq \nu(\mathcal{S}_0; P) \leq k.$$

Since the cardinality of P' is equal to $n - 1$, the induction hypothesis thus implies that $\operatorname{card}(\mathcal{S}_0(P')) \leq \Psi_k(n - 1)$.

On the other hand, we claim that the set class $\mathcal{S}_1$ satisfies the upper bound $\nu(\mathcal{S}_1; P') \leq k - 1$. Suppose that $\mathcal{S}_1$ shatters some subset $Q' \subseteq P'$ of cardinality m; it suffices to show that $m \leq k - 1$. If $\mathcal{S}_1$ shatters such a set Q', then $\mathcal{S}$ would shatter the set $Q = Q' \cup \{x_1\} \subseteq P$. (This fact follows by construction of $\mathcal{S}_1$: for every binary vector in the set $\mathcal{S}_1(P)$, the set $\mathcal{S}(P)$ must contain a binary vector with the label for x_1 flipped; see the concrete example given above for an illustration.) Since $\nu(\mathcal{S}; P) \leq k$, it must be the case that $\operatorname{card}(Q) = m + 1 \leq k$, which implies that $\nu(\mathcal{S}_1; P') \leq k - 1$. Consequently, the induction hypothesis implies that $\operatorname{card}(\mathcal{S}_1(P')) \leq \Psi_{k-1}(n - 1)$.

Putting together the pieces, we have shown that

$$\operatorname{card}(\mathcal{S}(P)) \leq \Psi_k(n - 1) + \Psi_{k-1}(n - 1) \overset{\text{(i)}}{=} \Psi_k(n), \tag{4.29}$$

where the equality (i) follows from an elementary combinatorial argument (see Exercise 4.10). This completes the proof. □

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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

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

We leave the proof of this result as an exercise for the reader (Exercise 4.8).

Vector space structure

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

Proposition 4.20 (Finite-dimensional vector spaces) *Let $\mathcal{G}$ be a vector space of functions $g: \mathbb{R}^d \to \mathbb{R}$ with dimension $\dim(\mathcal{G}) < \infty$. Then the subgraph class $\mathcal{S}(\mathcal{G})$ 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 $\mathcal{S}(\mathcal{G})$. Fix an arbitrary collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathbb{R}^d$, and consider the linear map $L: \mathcal{G} \to \mathbb{R}^n$ given by $L(g) = (g(x_1), \ldots, g(x_n))$. 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 coordinate is positive, and then write

$$\sum_{\{i \mid \gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i \mid \gamma_i > 0\}} \gamma_i g(x_i) \qquad \text{for all } g \in \mathcal{G}. \tag{4.30}$$

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

Proceeding via proof by contradiction, suppose that there were to exist some $g \in \mathcal{G}$ such that the associated subgraph set $S_g = \{x \in \mathbb{R}^d \mid g(x) \leq 0\}$ included only the subset $\{x_i \mid \gamma_i \leq 0\}$. For such a function g, the right-hand side of equation (4.30) would be strictly positive while the left-hand side would be non-positive, which is a contradiction. We conclude that $\mathcal{S}(\mathcal{G})$ fails to shatter the set $\{x_1, \ldots, x_n\}$, as claimed. □

Let us illustrate the use of Proposition 4.20 with some examples:

Example 4.21 (Linear functions in $\mathbb{R}^d$) For a pair $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, define the function $f_{a,b}(x) := \langle a, x \rangle + b$, and consider the family $\mathscr{L}^d := \{f_{a,b} \mid (a, b) \in \mathbb{R}^d \times \mathbb{R}\}$ of all such linear functions. The associated subgraph class $\mathcal{S}(\mathscr{L}^d)$ corresponds to the collection of 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 $\mathscr{L}^d$ forms a vector space of dimension $d + 1$, we obtain as an immediate consequence of Proposition 4.20 that $\mathcal{S}(\mathscr{L}^d)$ 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 $\mathcal{S}(\mathscr{L}^1)$ corresponds to the collection of all left-sided or right-sided intervals—that is,

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

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 $\mathcal{S}(\mathscr{L}^1)$ that contains x_2 while excluding both x_1 and x_3. This calculation shows that $\nu(\mathcal{S}(\mathscr{L}^1)) = 2$, which matches the upper bound obtained from Proposition 4.20. More generally, it can be shown that the VC dimension of $\mathcal{S}(\mathscr{L}^d)$ is $d + 1$, so that Proposition 4.20 yields a sharp result in all dimensions. ♣

Example 4.22 (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 $\mathcal{S}^d_{\text{sphere}}$ 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 subgraph of the function $f_{a,b}$.

In order to leverage Proposition 4.20, we first define a feature map $\phi \colon \mathbb{R}^d \to \mathbb{R}^{d+2}$ via $\phi(x) := (1, x_1, \ldots, x_d, \|x\|_2^2)$, and then consider functions of the form

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

The family of functions $\{g_c, c \in \mathbb{R}^{d+1}\}$ is a vector space of dimension $d + 2$, and it contains the function class $\{f_{a,b}, (a, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$. Consequently, by applying Proposition 4.20 to this larger vector space, we conclude that $\nu(\mathcal{S}^d_{\text{sphere}}) \leq d + 2$. 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 4.13 for an in-depth exploration of the case $d = 2$. ♣

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

4.4 Bibliographic details and background

First, a technical remark regarding measurability: in general, the normed difference $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ need not be measurable, since the function class $\mathscr{F}$ 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. Otherwise, for a general function class, there are various ways of dealing with the issue of measurability, including the use of outer probability (cf. van der Vaart and Wellner (1996)). Here we instead adopt the following convention, suitable for defining expectations of any function ϕ of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$. For any finite class of functions $\mathscr{G}$ contained within $\mathscr{F}$, the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{G}}$ is well defined, so that it is sensible to define

$$\mathbb{E}[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}})] := \sup\{\mathbb{E}[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{G}})] \mid \mathscr{G} \subset \mathscr{F}, \ \mathscr{G} \text{ has finite cardinality}\}.$$

By using this definition, we can always think instead of expectations defined via suprema over finite sets.

Theorem 4.4 was originally proved by Glivenko (1933) for the continuous case, and by Cantelli (1933) in the general setting. The non-asymptotic form of the Glivenko–Cantelli theorem given in Corollary 4.15 can be sharpened substantially. For instance, Dvoretsky, Kiefer and Wolfowitz (1956) prove that there is a constant C independent of F and n such that

$$\mathbb{P}[\|\widehat{F}_n - F\|_\infty \geq \delta] \leq Ce^{-2n\delta^2} \qquad \text{for all } \delta \geq 0. \tag{4.31}$$

Massart (1990) establishes the sharpest possible result, with the leading constant $C = 2$.

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 (Milman and Schechtman, 1986; Pisier, 1989; Ledoux and Talagrand, 1991). Rademacher and Gaussian complexities have also been studied extensively in the specific context of uniform laws of large numbers and empirical risk minimization (e.g. van der Vaart and Wellner, 1996; Koltchinskii and Panchenko, 2000; Koltchinskii, 2001, 2006; Bartlett and Mendelson, 2002; Bartlett et al., 2005). In Chapter 5, we develop further connections between these two forms of complexity, and the related notion of metric entropy.

Exercise 5.4 is adapted from Problem 2.6.3 from van der Vaart and Wellner (1996). The proof of Proposition 4.20 is adapted from Pollard (1984), who credits it to Steele (1978) and Dudley (1978).

4.5 Exercises

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

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

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

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

(i) The mean functional $F \mapsto \int x\, dF(x)$.
(ii) The Cramér–von Mises functional $F \mapsto \int [F(x) - F_0(x)]^2\, dF_0(x)$.
(iii) The quantile functional $Q_\alpha(F) = \inf\{t \in \mathbb{R} \mid F(t) \geq \alpha\}$.

Exercise 4.2 (Failure of Glivenko–Cantelli) Recall from Example 4.7 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

$$\mathcal{R}_n(\mathcal{S}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{S \in \mathcal{S}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbb{I}_S[X_i]\right|\right] \geq \frac{1}{2}. \tag{4.32}$$

Discuss the connection to Theorem 4.10.

Exercise 4.3 (Maximum likelihood and uniform laws) Recall from Example 4.8 our discussion of empirical and population risks for maximum likelihood over a family of densities $\{p_\theta, \theta \in \Omega\}$.

(a) Compute the population risk $R(\theta, \theta^*) = \mathbb{E}_{\theta^*}\left[\log \frac{p_{\theta^*}(X)}{p_\theta(X)}\right]$ in the following cases:

(i) Bernoulli: $p_\theta(x) = \frac{e^{\theta x}}{1+e^{\theta x}}$ for $x \in \{0, 1\}$;
(ii) Poisson: $p_\theta(x) = \frac{e^{\theta x} e^{-\exp(\theta)}}{x!}$ for $x \in \{0, 1, 2, \ldots\}$;
(iii) multivariate Gaussian: p_θ is the density of an $\mathcal{N}(\theta, \mathbf{\Sigma})$ vector, where the covariance matrix $\mathbf{\Sigma}$ is known and fixed.

(b) For each of the above cases:

(i) Letting $\widehat{\theta}$ denote the maximum likelihood estimate, give an explicit expression for the excess risk $E(\widehat{\theta}, \theta^*) = R(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega} R(\theta, \theta^*)$.
(ii) Give an upper bound on the excess risk in terms of an appropriate Rademacher complexity.

Exercise 4.4 (Details of symmetrization argument)

(a) Prove that

$$\sup_{g \in \mathcal{G}} \mathbb{E}[g(X)] \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} |g(X)|\right].$$

Use this to complete the proof of inequality (4.17).

(b) Prove that for any convex and non-decreasing function Φ,

$$\sup_{g \in \mathcal{G}} \Phi(\mathbb{E}[|g(X)|]) \leq \mathbb{E}\left[\Phi\left(\sup_{g \in \mathcal{G}} |g(X)|\right)\right].$$

Use this bound to complete the proof of Proposition 4.11.

Exercise 4.5 (Necessity of vanishing Rademacher complexity) In this exercise, we work through the proof of Proposition 4.12.

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

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

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

(b) Use concentration results to complete the proof of Proposition 4.12.

Exercise 4.6 (Too many linear classifiers) Consider the function class

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

corresponding to the $\{-1, +1\}$-valued classification rules defined by linear functions in $\mathbb{R}^d$. Supposing that $d \geq n$, let $x_1^n = \{x_1, \ldots, x_n\}$ be a collection of vectors in $\mathbb{R}^d$ that are linearly independent. Show that the empirical Rademacher complexity satisfies

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

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

Exercise 4.7 (Basic properties of Rademacher complexity) Prove the following properties of the Rademacher complexity.

(a) $\mathcal{R}_n(\mathscr{F}) = \mathcal{R}_n(\operatorname{conv}(\mathscr{F}))$.
(b) Show that $\mathcal{R}_n(\mathscr{F} + \mathscr{G}) \leq \mathcal{R}_n(\mathscr{F}) + \mathcal{R}_n(\mathscr{G})$. Give an example to demonstrate that this bound cannot be improved in general.
(c) Given a fixed and uniformly bounded function g, show that

$$\mathcal{R}_n(\mathscr{F} + g) \leq \mathcal{R}_n(\mathscr{F}) + \frac{\|g\|_\infty}{\sqrt{n}}. \tag{4.33}$$

Exercise 4.8 (Operations on VC classes) Let $\mathcal{S}$ and $\mathcal{T}$ be two classes of sets with finite VC dimensions. Show that each of the following operations lead to a new set class also with finite VC dimension.

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

Exercise 4.9 Prove Lemma 4.14.

Exercise 4.10 Prove equality (i) in equation (4.29), namely that

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}.$$

Exercise 4.11 In this exercise, we complete the proof of Proposition 4.18.

(a) Prove inequality (ii) in (4.26).
(b) For $n \geq \nu$, prove the sharper upper bound $\operatorname{card}(\mathcal{S}(x_1^n)) \leq (\frac{en}{\nu})^\nu$. (*Hint:* You might find the result of Exercise 2.9 useful.)

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press

Exercise 4.12 (VC dimension of left-sided intervals) Consider the class of left-sided half-intervals in $\mathbb{R}^d$:

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

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

Exercise 4.13 (VC dimension of spheres) Consider the class of all spheres in $\mathbb{R}^2$—that is

$$\mathcal{S}^2_{\text{sphere}} := \{S_{a,b},\ (a,b) \in \mathbb{R}^2 \times \mathbb{R}_+\}, \tag{4.34}$$

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 $\mathcal{S}^2_{\text{sphere}}$ can shatter any subset of three points that are not collinear.
(b) Show that no subset of four points can be shattered, and conclude that the VC dimension is $\nu(\mathcal{S}^2_{\text{sphere}}) = 3$.

Exercise 4.14 (VC dimension of monotone Boolean conjunctions) For a positive integer $d \geq 2$, consider the function $h_S : \{0,1\}^d \to \{0,1\}$ of the form

$$h_s(x_1, \ldots, x_d) = \begin{cases} 1 & \text{if } x_j = 1 \text{ for all } j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

The set of all Boolean monomials $\mathfrak{B}_d$ consists of all such functions as S ranges over all subsets of $\{1, 2, \ldots, d\}$, along with the constant functions $h \equiv 0$ and $h \equiv 1$. Show that the VC dimension of $\mathfrak{B}_d$ is equal to d.

Exercise 4.15 (VC dimension of closed and convex sets) Show that the class $\mathcal{C}^d_{\text{cc}}$ 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.)

Exercise 4.16 (VC dimension of polygons) Compute the VC dimension of the set of all polygons in $\mathbb{R}^2$ with at most four vertices.

Exercise 4.17 (Infinite VC dimension) For a scalar $t \in \mathbb{R}$, consider the function $f_t(x) = \text{sign}(\sin(tx))$. Prove that the function class $\{f_t : [-1,1] \to \mathbb{R} \mid t \in \mathbb{R}\}$ has infinite VC dimension. (*Note:* This shows that VC dimension is *not* equivalent to the number of parameters in a function class.)

https://doi.org/10.1017/9781108627771.004 Published online by Cambridge University Press