3 Rademacher Complexity and VC-Dimension

The hypothesis sets typically used in machine learning are infinite. But the sample complexity bounds of the previous chapter are uninformative when dealing with infinite hypothesis sets. One could ask whether efficient learning from a finite sample is even possible when the hypothesis set H is infinite. Our analysis of the family of axis-aligned rectangles (Example 2.1) indicates that this is indeed possible at least in some cases, since we proved that that infinite concept class was PAC-learnable. Our goal in this chapter will be to generalize that result and derive general learning guarantees for infinite hypothesis sets.

A general idea for doing so consists of reducing the infinite case to the analysis of finite sets of hypotheses and then proceed as in the previous chapter. There are different techniques for that reduction, each relying on a different notion of complexity for the family of hypotheses. The first complexity notion we will use is that of *Rademacher complexity*. This will help us derive learning guarantees using relatively simple proofs based on McDiarmid's inequality, while obtaining high-quality bounds, including data-dependent ones, which we will frequently make use of in future chapters. However, the computation of the empirical Rademacher complexity is NP-hard for some hypothesis sets. Thus, we subsequently introduce two other purely combinatorial notions, the *growth function* and the *VC-dimension*. We first relate the Rademacher complexity to the growth function and then bound the growth function in terms of the VC-dimension. The VC-dimension is often easier to bound or estimate. We will review a series of examples showing how to compute or bound it, then relate the growth function and the VC-dimensions. This leads to generalization bounds based on the VC-dimension. Finally, we present lower bounds based on the VC-dimension both in the realizable and non-realizable cases, which will demonstrate the critical role of this notion in learning.

3.1 Rademacher complexity

We will continue to use H to denote a hypothesis set as in the previous chapters, and h an element of H. Many of the results of this section are general and hold for an arbitrary loss function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. To each $h\colon \mathcal{X} \to \mathcal{Y}$, we can associate a function g that maps $(x, y) \in \mathcal{X} \times \mathcal{Y}$ to $L(h(x), y)$ without explicitly describing the specific loss L used. In what follows G will generally be interpreted as *the family of loss functions associated to* H.

The Rademacher complexity captures the richness of a family of functions by measuring the degree to which a hypothesis set can fit random noise. The following states the formal definitions of the empirical and average Rademacher complexity.

Definition 3.1 **Empirical Rademacher complexity**
Let G be a family of functions mapping from Z to $[a, b]$ and $S = (z_1, \ldots, z_m)$ a fixed sample of size m with elements in Z. Then, the empirical Rademacher complexity *of G with respect to the sample S is defined as:*

$$\widehat{\mathfrak{R}}_S(G) = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right], \tag{3.1}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$, with σ_is independent uniform random variables taking values in $\{-1, +1\}$.[1] *The random variables σ_i are called* Rademacher variables.

Let $\mathbf{g}_S$ denote the vector of values taken by function g over the sample S: $\mathbf{g}_S = (g(z_1), \ldots, g(z_m))^\top$. Then, the empirical Rademacher complexity can be rewritten as

$$\widehat{\mathfrak{R}}_S(G) = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{\boldsymbol{\sigma} \cdot \mathbf{g}_S}{m}\right].$$

The inner product $\boldsymbol{\sigma} \cdot \mathbf{g}_S$ measures the correlation of $\mathbf{g}_S$ with the vector of random noise $\boldsymbol{\sigma}$. The supremum $\sup_{g \in G} \frac{\boldsymbol{\sigma} \cdot \mathbf{g}_S}{m}$ is a measure of how well the function class G correlates with $\boldsymbol{\sigma}$ over the sample S. Thus, the empirical Rademacher complexity measures on average how well the function class G correlates with random noise on S. This describes the richness of the family G: richer or more complex families G can generate more vectors $\mathbf{g}_S$ and thus better correlate with random noise, on average.

1. We assume implicitly that the supremum over the family G in this definition is measurable and in general will adopt the same assumption throughout this book for other suprema over a class of functions. This assumption does not hold for arbitrary function classes but it is valid for the hypotheses sets typically considered in practice in machine learning, and the instances discussed in this book.

Definition 3.2 **Rademacher complexity**
Let D denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity *of G is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to D:*

$$\mathfrak{R}_m(G) = \mathop{\mathrm{E}}_{S\sim D^m}[\widehat{\mathfrak{R}}_S(G)]. \tag{3.2}$$

We are now ready to present our first generalization bounds based on Rademacher complexity.

Theorem 3.1
Let G be a family of functions mapping from Z to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $g \in G$:

$$\mathrm{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.3}$$

$$\textit{and} \quad \mathrm{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\,. \tag{3.4}$$

Proof For any sample $S = (z_1, \ldots, z_m)$ and any $g \in G$, we denote by $\widehat{\mathrm{E}}_S[g]$ the empirical average of g over S: $\widehat{\mathrm{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$. The proof consists of applying McDiarmid's inequality to function Φ defined for any sample S by

$$\Phi(S) = \sup_{g\in G} \mathrm{E}[g] - \widehat{\mathrm{E}}_S[g]. \tag{3.5}$$

Let S and S' be two samples differing by exactly one point, say z_m in S and z'_m in S'. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\Phi(S') - \Phi(S) \leq \sup_{g\in G} \widehat{\mathrm{E}}_S[g] - \widehat{\mathrm{E}}_{S'}[g] = \sup_{g\in G} \frac{g(z_m) - g(z'_m)}{m} \leq \frac{1}{m}. \tag{3.6}$$

Similarly, we can obtain $\Phi(S) - \Phi(S') \leq 1/m$, thus $|\Phi(S) - \Phi(S')| \leq 1/m$. Then, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta/2$, the following holds:

$$\Phi(S) \leq \mathop{\mathrm{E}}_{S}[\Phi(S)] + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{3.7}$$

We next bound the expectation of the right-hand side as follows:

$$\begin{aligned}
\mathop{\mathrm{E}}_{S}[\Phi(S)] &= \mathop{\mathrm{E}}_{S}\Big[\sup_{g\in H} \mathrm{E}[g] - \widehat{\mathrm{E}}_S(g)\Big] \\
&= \mathop{\mathrm{E}}_{S}\Big[\sup_{g\in H} \mathop{\mathrm{E}}_{S'}\big[\widehat{\mathrm{E}}_{S'}(g) - \widehat{\mathrm{E}}_S(g)\big]\Big] && (3.8)\\
&\leq \mathop{\mathrm{E}}_{S,S'}\Big[\sup_{g\in H} \widehat{\mathrm{E}}_{S'}(g) - \widehat{\mathrm{E}}_S(g)\Big] && (3.9)\\
&= \mathop{\mathrm{E}}_{S,S'}\Big[\sup_{g\in H} \frac{1}{m}\sum_{i=1}^{m}(g(z'_i) - g(z_i))\Big] && (3.10)\\
&= \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S,S'}\Big[\sup_{g\in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z'_i) - g(z_i))\Big] && (3.11)\\
&\leq \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S'}\Big[\sup_{g\in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z'_i)\Big] + \mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\Big[\sup_{g\in H} \frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i)\Big] && (3.12)\\
&= 2\mathop{\mathrm{E}}_{\boldsymbol{\sigma},S}\Big[\sup_{g\in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\Big] = 2\mathfrak{R}_m(G). && (3.13)
\end{aligned}$$

Equation 3.8 uses the fact that points in S' are sampled in an i.i.d. fashion and thus $\mathrm{E}[g] = \mathrm{E}_{S'}[\widehat{\mathrm{E}}_{S'}(g)]$, as in (2.3). Inequality 3.9 holds by Jensen's inequality and the convexity of the supremum function. In equation 3.11, we introduce Rademacher variables σ_is, that is uniformly distributed independent random variables taking values in $\{-1, +1\}$ as in definition 3.2. This does not change the expectation appearing in (3.10): when $\sigma_i = 1$, the associated summand remains unchanged; when $\sigma_i = -1$, the associated summand flips signs, which is equivalent to swapping z_i and z'_i between S and S'. Since we are taking the expectation over all possible S and S', this swap does not affect the overall expectation. We are simply changing the order of the summands within the expectation. (3.12) holds by the sub-additivity of the supremum function, that is the identity $\sup(U+V) \leq \sup(U)+\sup(V)$. Finally, (3.13) stems from the definition of Rademacher complexity and the fact that the variables σ_i and $-\sigma_i$ are distributed in the same way.

The reduction to $\mathfrak{R}_m(G)$ in equation 3.13 yields the bound in equation 3.3, using δ instead of $\delta/2$. To derive a bound in terms of $\widehat{\mathfrak{R}}_S(G)$, we observe that, by definition 3.2, changing one point in S changes $\widehat{\mathfrak{R}}_S(G)$ by at most $1/m$. Then, using again McDiarmid's inequality, with probability $1 - \delta/2$ the following holds:

$$\mathfrak{R}_m(G) \leq \widehat{\mathfrak{R}}_S(G) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{3.14}$$

Finally, we use the union bound to combine inequalities 3.7 and 3.14, which yields

with probability at least $1-\delta$:

$$\Phi(S) \leq 2\widehat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}, \tag{3.15}$$

which matches (3.4). ■

The following result relates the empirical Rademacher complexities of a hypothesis set H and to the family of loss functions G associated to H in the case of binary loss (zero-one loss).

Lemma 3.1
Let H be a family of functions taking values in $\{-1,+1\}$ and let G be the family of loss functions associated to H for the zero-one loss: $G = \{(x,y) \mapsto 1_{h(x)\neq y} \colon h \in H\}$. For any sample $S = ((x_1,y_1),\ldots,(x_m,y_m))$ of elements in $\mathcal{X} \times \{-1,+1\}$, let $S_{\mathcal{X}}$ denote its projection over $\mathcal{X}$: $S_{\mathcal{X}} = (x_1,\ldots,x_m)$. Then, the following relation holds between the empirical Rademacher complexities of G and H:

$$\widehat{\mathfrak{R}}_S(G) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(H). \tag{3.16}$$

Proof For any sample $S = ((x_1,y_1),\ldots,(x_m,y_m))$ of elements in $\mathcal{X} \times \{-1,+1\}$, by definition, the empirical Rademacher complexity of G can be written as:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(G) &= \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x_i)\neq y_i}\Big] \\
&= \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \tfrac{1-y_i h(x_i)}{2}\Big] \\
&= \frac{1}{2}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i y_i h(x_i)\Big] \\
&= \frac{1}{2}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\Big] = \frac{1}{2}\mathfrak{R}_{S_{\mathcal{X}}}(H),
\end{aligned}$$

where we used the fact that $1_{h(x_i)\neq y_i} = (1-y_i h(x_i))/2$ and the fact that for a fixed $y_i \in \{-1,+1\}$, σ_i and $-y_i\sigma_i$ are distributed in the same way. ■

Note that the lemma implies, by taking expectations, that for any $m \geq 1$, $\mathfrak{R}_m(G) = \frac{1}{2}\mathfrak{R}_m(H)$. These connections between the empirical and average Rademacher complexities can be used to derive generalization bounds for binary classification in terms of the Rademacher complexity of the hypothesis set H.

Theorem 3.2 **Rademacher complexity bounds – binary classification**
Let H be a family of functions taking values in $\{-1,+1\}$ and let D be the distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over

a sample S of size m drawn according to D, each of the following holds for any $h \in H$:

$$R(h) \leq \widehat{R}(h) + \mathfrak{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{3.17}$$

$$\textit{and} \quad R(h) \leq \widehat{R}(h) + \widehat{\mathfrak{R}}_S(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{3.18}$$

Proof The result follows immediately by theorem 3.1 and lemma 3.1. ■

The theorem provides two generalization bounds for binary classification based on the Rademacher complexity. Note that the second bound, (3.18), is data-dependent: the empirical Rademacher complexity $\widehat{\mathfrak{R}}_S(H)$ is a function of the specific sample S drawn. Thus, this bound could be particularly informative if we could compute $\widehat{\mathfrak{R}}_S(H)$. But, how can we compute the empirical Rademacher complexity? Using again the fact that σ_i and $-\sigma_i$ are distributed in the same way, we can write

$$\widehat{\mathfrak{R}}_S(H) = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)\Big] = -\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\inf_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\Big].$$

Now, for a fixed value of $\boldsymbol{\sigma}$, computing $\inf_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)$ is equivalent to an *empirical risk minimization* problem, which is known to be computationally hard for some hypothesis sets. Thus, in some cases, computing $\widehat{\mathfrak{R}}_S(H)$ could be computationally hard. In the next sections, we will relate the Rademacher complexity to combinatorial measures that are easier to compute.

3.2 Growth function

Here we will show how the Rademacher complexity can be bounded in terms of the *growth function*.

Definition 3.3 **Growth function**
The growth function $\Pi_H : \mathbb{N} \to \mathbb{N}$ *for a hypothesis set H is defined by:*

$$\forall m \in \mathbb{N},\ \Pi_H(m) = \max_{\{x_1,\ldots,x_m\}\subseteq X} \Big|\big\{\big(h(x_1),\ldots,h(x_m)\big) : h \in H\big\}\Big|. \tag{3.19}$$

Thus, $\Pi_H(m)$ is the maximum number of distinct ways in which m points can be classified using hypotheses in H. This provides another measure of the richness of the hypothesis set H. However, unlike the Rademacher complexity, this measure does not depend on the distribution, it is purely combinatorial.

To relate the Rademacher complexity to the growth function, we will use Massart's lemma.

Theorem 3.3 **Massart's lemma**
Let $A \subseteq \mathbb{R}^m$ be a finite set, with $r = \max_{\mathbf{x}\in A} \|\mathbf{x}\|_2$, then the following holds:

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{r\sqrt{2\log|A|}}{m}, \tag{3.20}$$

where σ_is are independent uniform random variables taking values in $\{-1,+1\}$ and $x_1,\ldots,x_m$ are the components of vector $\mathbf{x}$.

Proof For any $t>0$, using Jensen's inequality, rearranging terms, and bounding the supremum by a sum, we obtain:

$$\begin{aligned}\exp\left(t\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leq \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left(\exp\left[t\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right)\\ &= \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left(\sup_{x\in A}\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right) \leq \sum_{x\in A}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left(\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right).\end{aligned}$$

We next use the independence of the σ_is, then apply Hoeffding's lemma (lemma D.1), and use the definition of r to write:

$$\begin{aligned}\exp\left(t\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leq \sum_{x\in A}\Pi_{i=1}^{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}_i}\left(\exp\left[t\sigma_i x_i\right]\right)\\ &\leq \sum_{x\in A}\Pi_{i=1}^{m}\exp\left[\frac{t^2(2x_i)^2}{8}\right]\\ &= \sum_{x\in A}\exp\left[\frac{t^2}{2}\sum_{i=1}^{m}x_i^2\right] \leq \sum_{x\in A}\exp\left[\frac{t^2r^2}{2}\right] = |A|e^{\frac{t^2R^2}{2}}.\end{aligned}$$

Taking the log of both sides and dividing by t gives us:

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{\log|A|}{t}+\frac{tr^2}{2}. \tag{3.21}$$

If we choose $t=\frac{\sqrt{2\log|A|}}{r}$, which minimizes this upper bound, we get:

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq r\sqrt{2\log|A|}. \tag{3.22}$$

Dividing both sides by m leads to the statement of the lemma. ■

Using this result, we can now bound the Rademacher complexity in terms of the growth function.

Corollary 3.1
Let G be a family of functions taking values in $\{-1, +1\}$. Then the following holds:

$$\mathfrak{R}_m(G) \leq \sqrt{\frac{2 \log \Pi_G(m)}{m}}. \tag{3.23}$$

Proof For a fixed sample $S = (x_1, \ldots, x_m)$, we denote by $G_{|S}$ the set of vectors of function values $(g(x_1), \ldots, g(x_m))^\top$ where g is in G. Since $g \in G$ takes values in $\{-1, +1\}$, the norm of these vectors is bounded by $\sqrt{m}$. We can then apply Massart's lemma as follows:

$$\mathfrak{R}_m(G) = \mathop{\mathrm{E}}_{S}\left[\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{u \in G_{|S}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i u_i\right]\right] \leq \mathop{\mathrm{E}}_{S}\left[\frac{\sqrt{m}\sqrt{2 \log |G_{|S}|}}{m}\right].$$

By definition, $|G_{|S}|$ is bounded by the growth function, thus,

$$\mathfrak{R}_m(G) \leq \mathop{\mathrm{E}}_{S}\left[\frac{\sqrt{m}\sqrt{2 \log \Pi_G(m)}}{m}\right] = \sqrt{\frac{2 \log \Pi_G(m)}{m}},$$

which concludes the proof. ■

Combining the generalization bound (3.17) of theorem 3.2 with corollary 3.1 yields immediately the following generalization bound in terms of the growth function.

Corollary 3.2 **Growth function generalization bound**
Let H be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

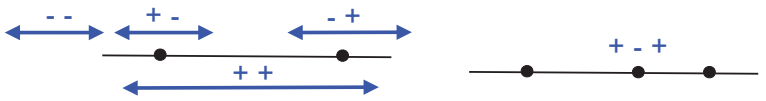
$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2 \log \Pi_H(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3.24}$$

Growth function bounds can be also derived directly (without using Rademacher complexity bounds first). The resulting bound is then the following:

$$\Pr\left[\left|R(h) - \widehat{R}(h)\right| > \epsilon\right] \leq 4\Pi_H(2m) \exp\left(-\frac{m\epsilon^2}{8}\right), \tag{3.25}$$

which only differs from (3.24) by constants.

The computation of the growth function may not be always convenient since, by definition, it requires computing $\Pi_H(m)$ for all $m \geq 1$. The next section introduces an alternative measure of the complexity of a hypothesis set H that is based instead on a single scalar, which will turn out to be in fact deeply related to the behavior of the growth function.



(a) (b)

Figure 3.1 VC-dimension of intervals on the real line. (a) Any two points can be shattered. (b) No sample of three points can be shattered as the $(+,-,+)$ labeling cannot be realized.

3.3 VC-dimension

Here, we introduce the notion of *VC-dimension* (Vapnik-Chervonenkis dimension). The VC-dimension is also a purely combinatorial notion but it is often easier to compute than the growth function (or the Rademacher Complexity). As we shall see, the VC-dimension is a key quantity in learning and is directly related to the growth function.

To define the VC-dimension of a hypothesis set H, we first introduce the concepts of *dichotomy* and that of *shattering*. Given a hypothesis set H, a dichotomy of a set S is one of the possible ways of labeling the points of S using a hypothesis in H. A set S of $m \geq 1$ points is said to be shattered by a hypothesis set H when H realizes all possible dichotomies of S, that is when $\Pi_H(m) = 2^m$.

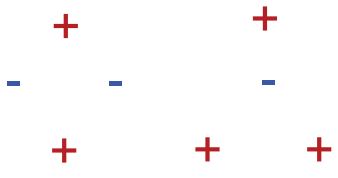
Definition 3.4 **VC-dimension**
The VC-dimension of a hypothesis set H is the size of the largest set that can be fully shattered by H:

$$VCdim(H) = \max\{m \colon \Pi_H(m) = 2^m\}. \tag{3.26}$$

Note that, by definition, if $\text{VCdim}(H) = d$, there exists a set of size d that can be fully shattered. But, this does not imply that all sets of size d or less are fully shattered, in fact, this is typically not the case.

To further illustrate this notion, we will examine a series of examples of hypothesis sets and will determine the VC-dimension in each case. To compute the VC-dimension we will typically show a lower bound for its value and then a matching upper bound. To give a lower bound d for $\text{VCdim}(H)$, it suffices to show that a set S of cardinality d can be shattered by H. To give an upper bound, we need to prove that no set S of cardinality $d+1$ can be shattered by H, which is typically more difficult.

Example 3.1 Intervals on the real line
Our first example involves the hypothesis class of intervals on the real line. It is clear that the VC-dimension is at least two, since all four dichotomies



(a) (b)

Figure 3.2 Unrealizable dichotomies for four points using hyperplanes in $\mathbb{R}^2$. (a) All four points lie on the convex hull. (b) Three points lie on the convex hull while the remaining point is interior.

$(+,+),(-,-),(+,-),(-,+)$ can be realized, as illustrated in figure 3.1(a). In contrast, by the definition of intervals, no set of three points can be shattered since the $(+,-,+)$ labeling cannot be realized. Hence, VCdim(intervals in $\mathbb{R}$) = 2.

Example 3.2 Hyperplanes

Consider the set of hyperplanes in $\mathbb{R}^2$. We first observe that any three non-collinear points in $\mathbb{R}^2$ can be shattered. To obtain the first three dichotomies, we choose a hyperplane that has two points on one side and the third point on the opposite side. To obtain the fourth dichotomy we have all three points on the same side of the hyperplane. The remaining four dichotomies are realized by simply switching signs. Next, we show that four points cannot be shattered by considering two cases: (i) the four points lie on the convex hull defined by the four points, and (ii) three of the four points lie on the convex hull and the remaining point is internal. In the first case, a positive labeling for one diagonal pair and a negative labeling for the other diagonal pair cannot be realized, as illustrated in figure 3.2(a). In the second case, a labeling which is positive for the points on the convex hull and negative for the interior point cannot be realized, as illustrated in figure 3.2(b). Hence, VCdim(hyperplanes in $\mathbb{R}^2$) = 3.

More generally in $\mathbb{R}^d$, we derive a lower bound by starting with a set of $d+1$ points in $\mathbb{R}^d$, setting x_0 to be the origin and defining x_i, for $i \in \{1,\ldots,d\}$, as the point whose ith coordinate is 1 and all others are 0. Let $y_0, y_1, \ldots, y_d \in \{-1,+1\}$ be an arbitrary set of labels for $x_0, x_1, \ldots, x_d$. Let w be the vector whose ith coordinate is y_i. Then the classifier defined by the hyperplane of equation $w \cdot x + \frac{y_0}{2} = 0$ shatters $x_0, x_1, \ldots, x_d$ since for any $i \in [0,d]$,

$$\operatorname{sgn}\left(w \cdot x_i + \frac{y_0}{2}\right) = \operatorname{sgn}\left(y_i + \frac{y_0}{2}\right) = y_i. \tag{3.27}$$

To obtain an upper bound, it suffices to show that no set of $d+2$ points can be shattered by halfspaces. To prove this, we will use the following general theorem.

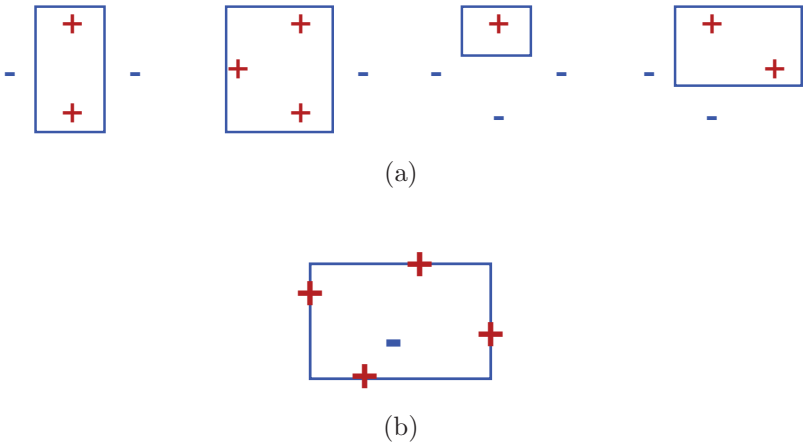


Figure 3.3 VC-dimension of axis-aligned rectangles. (a) Examples of realizable dichotomies for four points in a diamond pattern. (b) No sample of five points can be realized if the interior point and the remaining points have opposite labels.

Theorem 3.4 **Radon's theorem**
Any set X of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.

Proof Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{d+2}\} \subset \mathbb{R}^d$. The following is a system of $d+1$ linear equations in $\alpha_1, \ldots, \alpha_{d+2}$:

$$\sum_{i=1}^{d+2} \alpha_i \mathbf{x}_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{d+2} \alpha_i = 0, \tag{3.28}$$

since the first equality leads to d equations, one for each component. The number of unknowns, $d+2$, is larger than the number of equations, $d+1$, therefore the system admits a non-zero solution $\beta_1, \ldots, \beta_{d+2}$. Since $\sum_{i=1}^{d+2} \beta_i = 0$, both $I_1 = \{i \in [1, d+2] \colon \beta_i > 0\}$ and $I_2 = \{i \in [1, d+2] \colon \beta_i < 0\}$ are non-empty sets and $X_1 = \{\mathbf{x}_i \colon i \in I_1\}$ and $X_2 = \{\mathbf{x}_i \colon i \in I_2\}$ form a partition of X. By the last equation of (3.28), $\sum_{i \in I_1} \beta_i = -\sum_{i \in I_2} \beta_i$. Let $\beta = \sum_{i \in I_1} \beta_i$. Then, the first part of (3.28) implies

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} \mathbf{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} \mathbf{x}_i,$$

with $\sum_{i \in I_1} \frac{\beta_i}{\beta} = \sum_{i \in I_2} \frac{-\beta_i}{\beta} = 1$, and $\frac{\beta_i}{\beta} \geq 0$ for $i \in I_1$ and $\frac{-\beta_i}{\beta} \geq 0$ for $i \in I_2$. By definition of the convex hulls (B.4), this implies that $\sum_{i \in I_1} \frac{\beta_i}{\beta} \mathbf{x}_i$ belongs both to

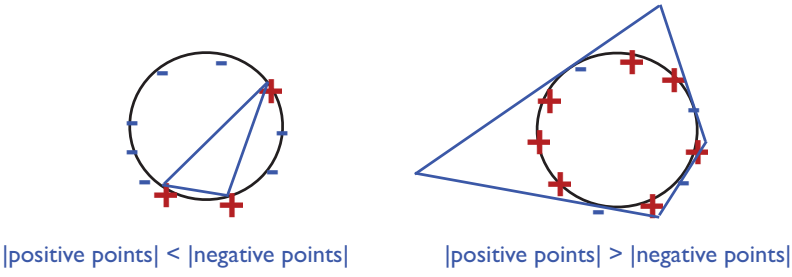


Figure 3.4 Convex d-gons in the plane can shatter $2d + 1$ points. (a) d-gon construction when there are more negative labels. (b) d-gon construction when there are more positive labels.

the convex hull of X_1 and to that of X_2. ■

Now, let X be a set of $d + 2$ points. By Radon's theorem, it can be partitioned into two sets X_1 and X_2 such that their convex hulls intersect. Observe that when two sets of points X_1 and X_2 are separated by a hyperplane, their convex hulls are also separated by that hyperplane. Thus, X_1 and X_2 cannot be separated by a hyperplane and X is not shattered. Combining our lower and upper bounds, we have proven that VCdim(hyperplanes in $\mathbb{R}^d$) $= d + 1$.

Example 3.3 Axis-aligned Rectangles

We first show that the VC-dimension is at least four, by considering four points in a diamond pattern. Then, it is clear that all 16 dichotomies can be realized, some of which are illustrated in figure 3.2(a). In contrast, for any set of five distinct points, if we construct the minimal axis-aligned rectangle containing these points, one of the five points is in the interior of this rectangle. Imagine that we assign a negative label to this interior point and a positive label to each of the remaining four points, as illustrated in figure 3.2(b). There is no axis-aligned rectangle that can realize this labeling. Hence, no set of five distinct points can be shattered and VCdim(axis-aligned rectangles) $= 4$.

Example 3.4 Convex Polygons

We focus on the class of convex d-gons in the plane. To get a lower bound, we show that any set of $2d + 1$ points can be fully shattered. To do this, we select $2d + 1$ points that lie on a circle, and for a particular labeling, if there are more negative than positive labels, then the points with the positive labels are used as the polygon's vertices, as in figure 3.4(a). Otherwise, the tangents of the negative points serve as the edges of the polygon, as shown in (3.4)(b). To derive an upper

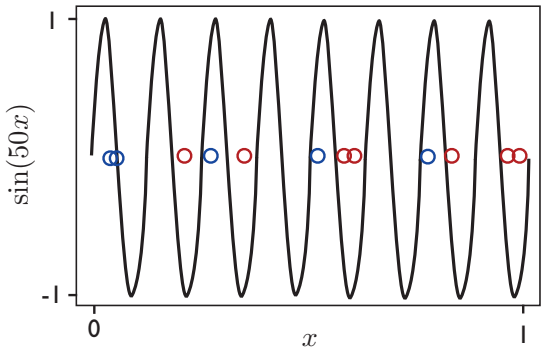


Figure 3.5 An example of a sine function (with $\omega = 50$) used for classification.

bound, it can be shown that choosing points on the circle maximizes the number of possible dichotomies, and thus VCdim(convex d-gons) $= 2d + 1$. Note also that VCdim(convex polygons) $= +\infty$.

Example 3.5 Sine Functions
The previous examples could suggest that the VC-dimension of H coincides with the number of free parameters defining H. For example, the number of parameters defining hyperplanes matches their VC-dimension. However, this does not hold in general. Several of the exercises in this chapter illustrate this fact. The following provides a striking example from this point of view. Consider the following family of sine functions: $\{t \mapsto \sin(\omega t) \colon \omega \in \mathbb{R}\}$. One instance of this function class is shown in figure 3.5. These sine functions can be used to classify the points on the real line: a point is labeled positively if it is above the curve, negatively otherwise. Although this family of sine function is defined via a single parameter, ω, it can be shown that VCdim(sine functions) $= +\infty$ (exercise 3.12).

The VC-dimension of many other hypothesis sets can be determined or upper-bounded in a similar way (see this chapter's exercises). In particular, the VC-dimension of any vector space of dimension $r < \infty$ can be shown to be at most r (exercise 3.11). The next result known as *Sauer's lemma* clarifies the connection between the notions of growth function and VC-dimension.

***Theorem 3.5* Sauer's lemma**
Let H be a hypothesis set with VCdim$(H) = d$. Then, for all $m \in \mathbb{N}$, the following inequality holds:

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}. \tag{3.29}$$

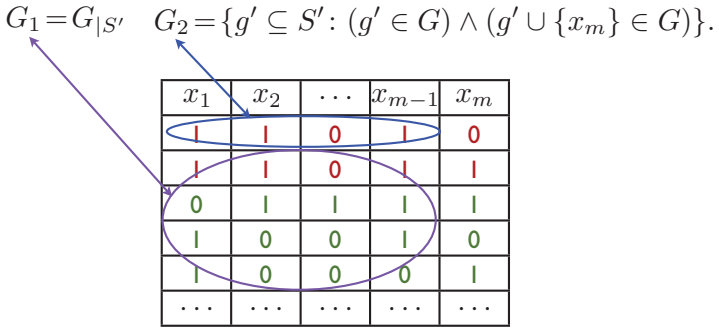


Figure 3.6 Illustration of how G_1 and G_2 are constructed in the proof of Sauer's lemma.

Proof The proof is by induction on $m + d$. The statement clearly holds for $m = 1$ and $d = 0$ or $d = 1$. Now, assume that it holds for $(m-1, d-1)$ and $(m-1, d)$. Fix a set $S = \{x_1, \ldots, x_m\}$ with $\Pi_H(m)$ dichotomies and let $G = H_{|S}$ be the set of concepts H induces by restriction to S.

Now consider the following families over $S' = \{x_1, \ldots, x_{m-1}\}$. We define $G_1 = G_{|S'}$ as the set of concepts H includes by restriction to S'. Next, by identifying each concept as the set of points (in S' or S) for which it is non-zero, we can define G_2 as

$$G_2 = \{g' \subseteq S' : (g' \in G) \wedge (g' \cup \{x_m\} \in G)\}.$$

Since $g' \subseteq S'$, $g' \in G$ means that without adding x_m it is a concept of G. Further, the constraint $g' \cup \{x_m\} \in G$ means that adding x_m to g' also makes it a concept of G. The construction of G_1 and G_2 is illustrated pictorially in figure 3.6. Given our definitions of G_1 and G_2, observe that $|G_1| + |G_2| = |G|$.

Since $\mathrm{VCdim}(G_1) \leq \mathrm{VCdim}(G) \leq d$, then by definition of the growth function and using the induction hypothesis,

$$|G_1| \leq \Pi_{G_1}(m-1) \leq \sum_{i=0}^{d} \binom{m-1}{i}.$$

Further, by definition of G_2, if a set $Z \subseteq S'$ is shattered by G_2, then the set $Z \cup \{x_m\}$ is shattered by G. Hence,

$$\mathrm{VCdim}(G_2) \leq \mathrm{VCdim}(G) - 1 = d - 1,$$

and by definition of the growth function and using the induction hypothesis,

$$|G_2| \leq \Pi_{G_2}(m-1) \leq \sum_{i=0}^{d-1} \binom{m-1}{i}.$$

Thus,

$$|G| = |G_1| + |G_2| \leq \sum_{i=0}^{d} \tbinom{m-1}{i} + \sum_{i=0}^{d-1} \tbinom{m-1}{i} = \sum_{i=0}^{d} \tbinom{m-1}{i} + \tbinom{m-1}{i-1} = \sum_{i=0}^{d} \tbinom{m}{i},$$

which completes the inductive proof. ■

The significance of Sauer's lemma can be seen by corollary 3.3, which remarkably shows that growth function only exhibits two types of behavior: either $\text{VCdim}(H) = d < +\infty$, in which case $\Pi_H(m) = O(m^d)$, or $\text{VCdim}(H) = +\infty$, in which case $\Pi_H(m) = 2^m$.

Corollary 3.3
Let H be a hypothesis set with VCdim$(H) = d$. Then for all $m \geq d$,

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d). \tag{3.30}$$

Proof The proof begins by using Sauer's lemma. The first inequality multiplies each summand by a factor that is greater than or equal to one since $m \geq d$, while the second inequality adds non-negative summands to the summation.

$$\begin{aligned}
\Pi_H(m) &\leq \sum_{i=0}^{d} \binom{m}{i} \\
&\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i \\
&= \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{m}{d}\right)^d e^d.
\end{aligned}$$

After simplifying the expression using the binomial theorem, the final inequality follows using the general identity $(1 - x) \leq e^{-x}$. ■

The explicit relationship just formulated between VC-dimension and the growth function combined with corollary 3.2 leads immediately to the following generaliza-

tion bounds based on the VC-dimension.

Corollary 3.4 **VC-dimension generalization bounds**
Let H be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H$:

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3.31}$$

Thus, the form of this generalization bound is

$$R(h) \leq \widehat{R}(h) + O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right), \tag{3.32}$$

which emphasizes the importance of the ratio m/d for generalization. The theorem provides another instance of Occam's razor principle where simplicity is measured in terms of smaller VC-dimension.

VC-dimension bounds can be derived directly without using an intermediate Rademacher complexity bound, as for (3.25): combining Sauer's lemma with (3.25) leads to the following high-probability bound

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}{m}},$$

which has the general form of (3.32). The log factor plays only a minor role in these bounds. A finer analysis can be used in fact to eliminate that factor.

3.4 Lower bounds

In the previous section, we presented several upper bounds on the generalization error. In contrast, this section provides lower bounds on the generalization error of any learning algorithm in terms of the VC-dimension of the hypothesis set used.

These lower bounds are shown by finding for any algorithm a 'bad' distribution. Since the learning algorithm is arbitrary, it will be difficult to specify that particular distribution. Instead, it suffices to prove its existence non-constructively. At a high level, the proof technique used to achieve this is the *probabilistic method* of Paul Erdös. In the context of the following proofs, first a lower bound is given on the expected error over the parameters defining the distributions. From that, the lower bound is shown to hold for at least one set of parameters, that is one distribution.

Theorem 3.6 **Lower bound, realizable case**
Let H be a hypothesis set with VC-dimension $d > 1$. Then, for any learning algorithm $\mathcal{A}$, there exist a distribution D over $\mathcal{X}$ and a target function $f \in H$ such that

$$\Pr_{S \sim D^m}\left[R_D(h_S, f) > \frac{d-1}{32m}\right] \geq 1/100. \tag{3.33}$$

Proof Let $X = \{x_0, x_1, \ldots, x_{d-1}\} \subseteq \mathcal{X}$ be a set that is fully shattered by H. For any $\epsilon > 0$, we choose D such that its support is reduced to X and so that one point (x_0) has very high probability ($1 - \epsilon$), with the rest of the probability mass distributed uniformly among the other points:

$$\Pr_D[x_0] = 1 - 8\epsilon \quad \text{and} \quad \forall i \in [1, d-1], \Pr_D[x_i] = \frac{8\epsilon}{d-1}. \tag{3.34}$$

With this definition, most samples would contain x_0 and, since X is fully shattered, $\mathcal{A}$ can essentially do no better than tossing a coin when determining the label of a point x_i not falling in the training set.

We assume without loss of generality that $\mathcal{A}$ makes no error on x_0. For a sample S, we let $\overline{S}$ denote the set of its elements falling in $\{x_1, \ldots, x_{d-1}\}$, and let $\mathcal{S}$ be the set of samples S of size m such that $|\overline{S}| \leq (d-1)/2$. Now, fix a sample $S \in \mathcal{S}$, and consider the uniform distribution U over all labelings $f\colon X \to \{0, 1\}$, which are all in H since the set is shattered. Then, the following lower bound holds:

$$\begin{aligned}
\mathop{\mathrm{E}}_{f \sim U}[R_D(h_S, f)] &= \sum_f \sum_{x \in X} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
&\geq \sum_f \sum_{x \notin \overline{S}} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
&= \sum_{x \notin \overline{S}} \Big(\sum_f 1_{h(x) \neq f(x)} \Pr[f] \Big) \Pr[x] \\
&= \frac{1}{2} \sum_{x \notin \overline{S}} \Pr[x] \geq \frac{1}{2} \frac{d-1}{2} \frac{8\epsilon}{d-1} = 2\epsilon.
\end{aligned} \tag{3.35}$$

The first lower bound holds because we remove non-negative terms from the summation when we only consider $x \notin \overline{S}$ instead of all x in X. After rearranging terms, the subsequent equality holds since we are taking an expectation over $f \in H$ with uniform weight on each f and H shatters X. The final lower bound holds due to the definitions of D and $\overline{S}$, the latter which implies that $|X - \overline{S}| \geq (d-1)/2$.

Since (3.35) holds for all $S \in \mathcal{S}$, it also holds in expectation over all $S \in \mathcal{S}$: $\mathrm{E}_{S \in \mathcal{S}}\left[\mathrm{E}_{f \sim U}[R_D(h_S, f)]\right] \geq 2\epsilon$. By Fubini's theorem, the expectations can be

permuted, thus,

$$\mathop{\mathrm{E}}_{f\sim U}\Big[\mathop{\mathrm{E}}_{S\in\mathcal{S}}[R_D(h_S,f)]\Big] \geq 2\epsilon. \tag{3.36}$$

This implies that $\mathrm{E}_{S\in\mathcal{S}}[R_D(h_S,f_0)] \geq 2\epsilon$ for at least one labeling $f_0 \in H$. Decomposing this expectation into two parts and using $R_D(h_S,f_0) \leq \Pr_D[X - \{x_0\}]$, we obtain:

$$\begin{aligned}\mathop{\mathrm{E}}_{S\in\mathcal{S}}[R_D(h_S,f_0)] &= \sum_{S:R_D(h_S,f_0)\geq\epsilon} R_D(h_S,f_0)\Pr[R_D(h_S,f_0)] + \sum_{S:R_D(h_S,f_0)<\epsilon} R_D(h_S,f_0)\Pr[R_D(h_S,f_0)]\\ &\leq \Pr_D[X-\{x_0\}]\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon] + \epsilon\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)<\epsilon]\\ &\leq 8\epsilon\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon] + \epsilon\big(1-\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon]\big).\end{aligned}$$

Collecting terms in $\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon]$ yields

$$\Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon] \geq \frac{1}{7\epsilon}(2\epsilon-\epsilon) = \frac{1}{7}. \tag{3.37}$$

Thus, the probability over all samples S (not necessarily in $\mathcal{S}$) can be lower bounded as

$$\Pr_S[R_D(h_S,f_0)\geq\epsilon] \geq \Pr_{S\in\mathcal{S}}[R_D(h_S,f_0)\geq\epsilon]\Pr[\mathcal{S}] \geq \frac{1}{7}\Pr[\mathcal{S}]. \tag{3.38}$$

This leads us to find a lower bound for $\Pr[\mathcal{S}]$. The probability that more than $(d-1)/2$ points are drawn in a sample of size m verifies the Chernoff bound for any $\gamma > 0$:

$$1-\Pr[\mathcal{S}] = \Pr[S_m \geq 8\epsilon m(1+\gamma)] \leq e^{-8\epsilon m\frac{\gamma^2}{3}}. \tag{3.39}$$

Therefore, for $\epsilon = (d-1)/(32m)$ and $\gamma = 1$,

$$\Pr[S_m \geq \tfrac{d-1}{2}] \leq e^{-(d-1)/12} \leq e^{-1/12} \leq 1-7\delta, \tag{3.40}$$

for $\delta \leq .01$. Thus $\Pr[\mathcal{S}] \geq 7\delta$ and $\Pr_S[R_D(h_S,f_0)\geq\epsilon] \geq \delta$. ■

The theorem shows that for any algorithm $\mathcal{A}$, there exists a 'bad' distribution over $\mathcal{X}$ and a target function f for which the error of the hypothesis returned by $\mathcal{A}$ is $\Omega(\frac{d}{m})$ with some constant probability. This further demonstrates the key role played by the VC-dimension in learning. The result implies in particular that PAC-learning in the non-realizable case is not possible when the VC-dimension is infinite.

Note that the proof shows a stronger result than the statement of the theorem: the distribution D is selected independently of the algorithm $\mathcal{A}$. We now present a theorem giving a lower bound in the non-realizable case. The following two lemmas will be needed for the proof.

Lemma 3.2
Let α be a uniformly distributed random variable taking values in $\{\alpha_-, \alpha_+\}$, where $\alpha_- = \frac{1}{2} - \frac{\epsilon}{2}$ and $\alpha_+ = \frac{1}{2} + \frac{\epsilon}{2}$, and let S be a sample of $m \geq 1$ random variables $X_1, \ldots, X_m$ taking values in $\{0,1\}$ and drawn i.i.d. according to the distribution D_α defined by $\Pr_{D_\alpha}[X = 1] = \alpha$. Let h be a function from $\mathcal{X}^m$ to $\{\alpha_-, \alpha_+\}$, then the following holds:

$$\mathop{\mathrm{E}}_{\alpha}\Big[\Pr_{S\sim D_\alpha^m}[h(S) \neq \alpha]\Big] \geq \Phi(2\lceil m/2\rceil, \epsilon), \tag{3.41}$$

where $\Phi(m,\epsilon) = \frac{1}{4}\Big(1 - \sqrt{1 - \exp\big(-\frac{m\epsilon^2}{1-\epsilon^2}\big)}\Big)$ for all m and ϵ.

Proof The lemma can be interpreted in terms of an experiment with two coins with biases α_- and α_+. It implies that for a discriminant rule $h(S)$ based on a sample S drawn from D_{α_-} or D_{α_+}, to determine which coin was tossed, the sample size m must be at least $\Omega(1/\epsilon^2)$. The proof is left as an exercise (exercise 3.19). ■

We will make use of the fact that for any fixed ϵ the function $m \mapsto \Phi(m,x)$ is convex, which is not hard to establish.

Lemma 3.3
Let Z be a random variable taking values in $[0,1]$. Then, for any $\gamma \in [0,1)$,

$$\Pr[z > \gamma] \geq \frac{\mathrm{E}[Z] - \gamma}{1 - \gamma} > \mathrm{E}[Z] - \gamma. \tag{3.42}$$

Proof Since the values taken by Z are in $[0,1]$,

$$\begin{aligned}
\mathrm{E}[Z] &= \sum_{z\leq\gamma} \Pr[Z = z]z + \sum_{z>\gamma} \Pr[Z = z]z \\
&\leq \sum_{z\leq\gamma} \Pr[Z = z]\gamma + \sum_{z>\gamma} \Pr[Z = z] \\
&= \gamma \Pr[Z \leq \gamma] + \Pr[Z > \gamma] \\
&= \gamma(1 - \Pr[Z > \gamma]) + \Pr[Z > \gamma] \\
&= (1-\gamma)\Pr[Z > \gamma] + \gamma,
\end{aligned}$$

which concludes the proof. ■

Theorem 3.7 **Lower bound, non-realizable case**
Let H be a hypothesis set with VC-dimension $d > 1$. Then, for any learning algorithm $\mathcal{A}$, there exists a distribution D over $\mathcal{X} \times \{0,1\}$ such that:

$$\Pr_{S\sim D^m}\left[R_D(h_S) - \inf_{h\in H} R_D(h) > \sqrt{\frac{d}{320m}}\right] \geq 1/64. \tag{3.43}$$

Equivalently, for any learning algorithm, the sample complexity verifies

$$m \geq \frac{d}{320\epsilon^2}. \tag{3.44}$$

Proof Let $X = \{x_1, x_1, \ldots, x_d\} \subseteq \mathcal{X}$ be a set fully shattered by H. For any $\alpha \in [0,1]$ and any vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)^\top \in \{-1,+1\}^d$, we define a distribution $D_{\boldsymbol{\sigma}}$ with support $X \times \{0,1\}$ as follows:

$$\forall i \in [1,d], \quad \Pr_{D_{\boldsymbol{\sigma}}}[(x_i,1)] = \frac{1}{d}\Big(\frac{1}{2} + \frac{\sigma_i \alpha}{2}\Big). \tag{3.45}$$

Thus, the label of each point x_i, $i \in [1,d]$, follows the distribution $\Pr_{D_{\boldsymbol{\sigma}}}[\cdot|x_i]$, that of a biased coin where the bias is determined by the sign of σ_i and the magnitude of α. To determine the most likely label of each point x_i, the learning algorithm will therefore need to estimate $\Pr_{D_{\boldsymbol{\sigma}}}[1|x_i]$ with an accuracy better than α. To make this further difficult, α and $\boldsymbol{\sigma}$ will be selected based on the algorithm, requiring, as in lemma 3.2, $\Omega(1/\alpha^2)$ instances of each point x_i in the training sample.

Clearly, the Bayes classifier $h^*_{D_{\boldsymbol{\sigma}}}$ is defined by $h^*_{D_{\boldsymbol{\sigma}}}(x_i) = \operatorname{argmax}_{y \in \{0,1\}} \Pr[y|x_i] = 1_{\sigma_i > 0}$ for all $i \in [1,d]$. $h^*_{D_{\boldsymbol{\sigma}}}$ is in H since X is fully shattered. For all $h \in H$,

$$R_{D_{\boldsymbol{\sigma}}}(h) - R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}}) = \frac{1}{d}\sum_{x \in X}\Big(\frac{\alpha}{2} + \frac{\alpha}{2}\Big) 1_{h(x) \neq h^*_{D_{\boldsymbol{\sigma}}}(x)} = \frac{\alpha}{d}\sum_{x \in X} 1_{h(x) \neq h^*_{D_{\boldsymbol{\sigma}}}(x)}. \tag{3.46}$$

Let h_S denote the hypothesis returned by the learning algorithm $\mathcal{A}$ after receiving a labeled sample S drawn according to $D_{\boldsymbol{\sigma}}$. We will denote by $|S|_x$ the number of occurrences of a point x in S. Let U denote the uniform distribution over $\{-1,+1\}^d$.

Then, in view of (3.46), the following holds:

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{\substack{\boldsymbol{\sigma}\sim U\\ S\sim D^m_{\boldsymbol{\sigma}}}}\left[\frac{1}{\alpha}\big[R_{D_{\boldsymbol{\sigma}}}(h_S)-R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}})\big]\right]\\
&\quad=\frac{1}{d}\sum_{x\in X}\mathop{\mathrm{E}}_{\substack{\boldsymbol{\sigma}\sim U\\ S\sim D^m_{\boldsymbol{\sigma}}}}\left[1_{h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)}\right]\\
&\quad=\frac{1}{d}\sum_{x\in X}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}\sim U}\left[\Pr_{S\sim D^m_{\boldsymbol{\sigma}}}\big[h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)\big]\right]\\
&\quad=\frac{1}{d}\sum_{x\in X}\sum_{n=0}^{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}\sim U}\left[\Pr_{S\sim D^m_{\boldsymbol{\sigma}}}\big[h_S(x)\neq h^*_{D_{\boldsymbol{\sigma}}}(x)\,\big|\,|S|_x=n\big]\Pr[|S|_x=n]\right]\\
&\quad\geq\frac{1}{d}\sum_{x\in X}\sum_{n=0}^{m}\Phi(n+1,\alpha)\Pr[|S|_x=n] && \text{(lemma 3.2)}\\
&\quad\geq\frac{1}{d}\sum_{x\in X}\Phi(m/d+1,\alpha) && \text{(convexity of } \Phi(\cdot,\alpha) \text{ and Jensen's ineq.)}\\
&\quad=\Phi(m/d+1,\alpha).
\end{aligned}
$$

Since the expectation over $\boldsymbol{\sigma}$ is lower-bounded by $\Phi(m/d+1,\alpha)$, there must exist some $\boldsymbol{\sigma}\in\{-1,+1\}^d$ for which

$$
\mathop{\mathrm{E}}_{S\sim D^m_{\boldsymbol{\sigma}}}\left[\frac{1}{\alpha}\big[R_{D_{\boldsymbol{\sigma}}}(h_S)-R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}})\big]\right]>\Phi(m/d+1,\alpha). \tag{3.47}
$$

Then, by lemma 3.3, for that $\boldsymbol{\sigma}$, for any $\gamma\in[0,1]$,

$$
\Pr_{S\sim D^m_{\boldsymbol{\sigma}}}\left[\frac{1}{\alpha}\big[R_{D_{\boldsymbol{\sigma}}}(h_S)-R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}})\big]>\gamma u\right]>(1-\gamma)u, \tag{3.48}
$$

where $u=\Phi(m/d+1,\alpha)$. Selecting δ and ϵ such that $\delta\leq(1-\gamma)u$ and $\epsilon\leq\gamma\alpha u$ gives

$$
\Pr_{S\sim D^m_{\boldsymbol{\sigma}}}\big[R_{D_{\boldsymbol{\sigma}}}(h_S)-R_{D_{\boldsymbol{\sigma}}}(h^*_{D_{\boldsymbol{\sigma}}})>\epsilon\big]>\delta. \tag{3.49}
$$

To satisfy the inequalities defining ϵ and δ, let $\gamma = 1 - 8\delta$. Then,

$$\delta \leq (1-\gamma)u \iff u \geq \frac{1}{8} \tag{3.50}$$

$$\iff \frac{1}{4}\left(1 - \sqrt{1 - \exp\left(-\frac{(m/d+1)\alpha^2}{1-\alpha^2}\right)}\right) \geq \frac{1}{8} \tag{3.51}$$

$$\iff \frac{(m/d+1)\alpha^2}{1-\alpha^2} \leq \log\frac{4}{3} \tag{3.52}$$

$$\iff \frac{m}{d} \leq (\frac{1}{\alpha^2} - 1)\log\frac{4}{3} - 1. \tag{3.53}$$

Selecting $\alpha = 8\epsilon/(1-8\delta)$ gives $\epsilon = \gamma\alpha/8$ and the condition

$$\frac{m}{d} \leq \left(\frac{(1-8\delta)^2}{64\epsilon^2} - 1\right)\log\frac{4}{3} - 1. \tag{3.54}$$

Let $f(1/\epsilon^2)$ denote the right-hand side. We are seeking a sufficient condition of the form $m/d \leq \omega/\epsilon^2$. Since $\epsilon \leq 1/64$, to ensure that $\omega/\epsilon^2 \leq f(1/\epsilon^2)$, it suffices to impose $\omega/(1/64)^2 = f(1/(1/64)^2)$. This condition gives

$$\omega = (7/64)^2\log(4/3) - (1/64)^2(\log(4/3)+1) \approx .003127 \geq 1/320 = .003125.$$

Thus, $\epsilon^2 \leq \frac{1}{320(m/d)}$ is sufficient to ensure the inequalities. ■

The theorem shows that for any algorithm $\mathcal{A}$, in the non-realizable case, there exists a 'bad' distribution over $\mathcal{X} \times \{0,1\}$ such that the error of the hypothesis returned by $\mathcal{A}$ is $\Omega\left(\sqrt{\frac{d}{m}}\right)$ with some constant probability. The VC-dimension appears as a critical quantity in learning in this general setting as well. In particular, with an infinite VC-dimension, agnostic PAC-learning is not possible.

3.5 Chapter notes

The use of Rademacher complexity for deriving generalization bounds in learning was first advocated by Koltchinskii [2001], Koltchinskii and Panchenko [2000], and Bartlett, Boucheron, and Lugosi [2002a], see also [Koltchinskii and Panchenko, 2002, Bartlett and Mendelson, 2002]. Bartlett, Bousquet, and Mendelson [2002b] introduced the notion of *local Rademacher complexity*, that is the Rademacher complexity restricted to a subset of the hypothesis set limited by a bound on the variance. This can be used to derive better guarantees under some regularity assumptions about the noise.

Theorem 3.3 is due to Massart [2000]. The notion of VC-dimension was introduced by Vapnik and Chervonenkis [1971] and has been since extensively studied [Vapnik,

2006, Vapnik and Chervonenkis, 1974, Blumer et al., 1989, Assouad, 1983, Dudley, 1999]. In addition to the key role it plays in machine learning, the VC-dimension is also widely used in a variety of other areas of computer science and mathematics (e.g., see Shelah [1972], Chazelle [2000]). Theorem 3.5 is known as *Sauer's lemma* in the learning community, however the result was first given by Vapnik and Chervonenkis [1971] (in a somewhat different version) and later independently by Sauer [1972] and Shelah [1972].

In the realizable case, lower bounds for the expected error in terms of the VC-dimension were given by Vapnik and Chervonenkis [1974] and Haussler et al. [1988]. Later, a lower bound for the probability of error such as that of theorem 3.6 was given by Blumer et al. [1989]. Theorem 3.6 and its proof, which improves upon this previous result, are due to Ehrenfeucht, Haussler, Kearns, and Valiant [1988]. Devroye and Lugosi [1995] gave slightly tighter bounds for the same problem with a more complex expression. Theorem 3.7 giving a lower bound in the non-realizable case and the proof presented are due to Anthony and Bartlett [1999]. For other examples of application of the probabilistic method demonstrating its full power, consult the reference book of Alon and Spencer [1992].

There are several other measures of the complexity of a family of functions used in machine learning, including *covering numbers*, *packing numbers*, and some other complexity measures discussed in chapter 10. A covering number $\mathcal{N}_p(G, \epsilon)$ is the minimal number of L_p balls of radius $\epsilon > 0$ needed to cover a family of loss functions G. A packing number $\mathcal{M}_p(G, \epsilon)$ is the maximum number of non-overlapping L_p balls of radius ϵ centered in G. The two notions are closely related, in particular it can be shown straightfowardly that $\mathcal{M}_p(G, 2\epsilon) \leq \mathcal{N}_p(G, \epsilon) \leq \mathcal{M}_p(G, \epsilon)$ for G and $\epsilon > 0$. Each complexity measure naturally induces a different reduction of infinite hypothesis sets to finite ones, thereby resulting in generalization bounds for infinite hypothesis sets. Exercise 3.22 illustrates the use of covering numbers for deriving generalization bounds using a very simple proof. There are also close relationships between these complexity measures: for example, by Dudley's theorem, the empirical Rademacher complexity can be bounded in terms of $\mathcal{N}_2(G, \epsilon)$ [Dudley, 1967, 1987] and the covering and packing numbers can be bounded in terms of the VC-dimension [Haussler, 1995]. See also [Ledoux and Talagrand, 1991, Alon et al., 1997, Anthony and Bartlett, 1999, Cucker and Smale, 2001, Vidyasagar, 1997] for a number of upper bounds on the covering number in terms of other complexity measures.

3.6 Exercises

3.1 Growth function of intervals in $\mathbb{R}$. Let H be the set of intervals in $\mathbb{R}$. The VC-dimension of H is 2. Compute its shattering coefficient $\Pi_H(m)$, $m \geq 0$. Compare

your result with the general bound for growth functions.

3.2 Lower bound on growth function. Prove that Sauer's lemma (theorem 3.5) is tight, i.e., for any set X of $m > d$ elements, show that there exists a hypothesis class H of VC-dimension d such that $\Pi_H(m) = \sum_{i=0}^{d} \binom{m}{i}$.

3.3 Singleton hypothesis class. Consider the trivial hypothesis set $H = \{h_0\}$.

(a) Show that $\mathfrak{R}_m(H) = 0$ for any $m > 0$.

(b) Use a similar construction to show that Massart's lemma (theorem 3.3) is tight.

3.4 Rademacher identities. Fix $m \geq 1$. Prove the following identities for any $\alpha \in \mathbb{R}$ and any two hypothesis sets H and H' of functions mapping from $\mathcal{X}$ to $\mathbb{R}$:

(a) $\mathfrak{R}_m(\alpha H) = |\alpha| \mathfrak{R}_m(H)$.

(b) $\mathfrak{R}_m(H + H') = \mathfrak{R}_m(H) + \mathfrak{R}_m(H')$.

(c) $\mathfrak{R}_m(\{\max(h, h') \colon h \in H, h' \in H'\})$,
where $\max(h, h')$ denotes the function $x \mapsto \max_{x \in \mathcal{X}}(h(x), h'(x))$ (*Hint*: you could use the identity $\max(a, b) = \frac{1}{2}[a + b + |a - b|]$ valid for all $a, b \in \mathbb{R}$ and Talagrand's contraction lemma (see lemma 4.2)).

3.5 Rademacher complexity. Professor Jesetoo claims to have found a better bound on the Rademacher complexity of any hypothesis set H of functions taking values in $\{-1, +1\}$, in terms of its VC-dimension $\mathrm{VCdim}(H)$. His bound is of the form $\mathfrak{R}_m(H) \leq O\big(\frac{\mathrm{VCdim}(H)}{m}\big)$. Can you show that Professor Jesetoo's claim cannot be correct? (*Hint*: consider a hypothesis set H reduced to just two simple functions.)

3.6 VC-dimension of union of k intervals. What is the VC-dimension of subsets of the real line formed by the union of k intervals?

3.7 VC-dimension of finite hypothesis sets. Show that the VC-dimension of a finite hypothesis set H is at most $\log_2 |H|$.

3.8 VC-dimension of subsets. What is the VC-dimension of the set of subsets I_α of the real line parameterized by a single parameter α: $I_\alpha = [\alpha, \alpha + 1] \cup [\alpha + 2, +\infty)$?

3.9 VC-dimension of closed balls in $\mathbb{R}^n$. Show that the VC-dimension of the set of all closed balls in $\mathbb{R}^n$, i.e., sets of the form $\{x \in \mathbb{R}^n \colon \|x - x_0\|^2 \leq r\}$ for some $x_0 \in \mathbb{R}^n$ and $r \geq 0$, is less than or equal to $n + 2$.

3.10 VC-dimension of ellipsoids. What is the VC-dimension of the set of all ellipsoids in $\mathbb{R}^n$?

3.11 VC-dimension of a vector space of real functions. Let F be a finite-dimensional vector space of real functions on $\mathbb{R}^n$, $\dim(F) = r < \infty$. Let H be the set of hypotheses:

$$H = \{\{x\colon f(x) \geq 0\}\colon f \in F\}.$$

Show that d, the VC-dimension of H, is finite and that $d \leq r$. (*Hint*: select an arbitrary set of $m = r + 1$ points and consider linear mapping $u\colon F \to \mathbb{R}^m$ defined by: $u(f) = (f(x_1), \ldots, f(x_m))$.)

3.12 VC-dimension of sine functions. Consider the hypothesis family of sine functions (Example 3.5): $\{x \to \sin(\omega x)\colon \omega \in \mathbb{R}\}$.

(a) Show that for any $x \in \mathbb{R}$ the points $x, 2x, 3x$ and $4x$ cannot be shattered by this family of sine functions.

(b) Show that the VC-dimension of the family of sine functions is infinite. (*Hint*: show that $\{2^{-m}\colon m \in \mathbb{N}\}$ can be fully shattered for any $m > 0$.)

3.13 VC-dimension of union of halfspaces. Determine the VC-dimension of the subsets of the real line formed by the union of k intervals.

3.14 VC-dimension of intersection of halfspaces. Consider the class C_k of convex intersections of k halfspaces. Give lower and upper bound estimates for $\mathrm{VCdim}(C_k)$.

3.15 VC-dimension of intersection concepts.

(a) Let C_1 and C_2 be two concept classes. Show that for any concept class $C = \{c_1 \cap c_2\colon c_1 \in C_1, c_2 \in C_2\}$,

$$\Pi_C(m) \leq \Pi_{C_1}(m)\,\Pi_{C_2}(m). \tag{3.55}$$

(b) Let C be a concept class with VC-dimension d and let C_s be the concept class formed by all intersections of s concepts from C, $s \geq 1$. Show that the VC-dimension of C_s is bounded by $2ds\log_2(3s)$. (*Hint*: show that $\log_2(3x) < 9x/(2e)$ for any $x \geq 2$.)

3.16 VC-dimension of union of concepts. Let A and B be two sets of functions mapping from X into $\{0, 1\}$, and assume that both A and B have finite VC-dimension, with $\mathrm{VCdim}(A) = d_A$ and $\mathrm{VCdim}(B) = d_B$. Let $C = A \cup B$ be the

union of A and B.

(a) Prove that for all m, $\Pi_C(m) \leq \Pi_A(m) + \Pi_B(m)$.

(b) Use Sauer's lemma to show that for $m \geq d_A + d_B + 2$, $\Pi_C(m) < 2^m$, and give a bound on the VC-dimension of C.

3.17 VC-dimension of symmetric difference of concepts. For two sets A and B, let $A \Delta B$ denote the symmetric difference of A and B, i.e., $A \Delta B = (A \cup B) - (A \cap B)$. Let H be a non-empty family of subsets of X with finite VC-dimension. Let A be an element of H and define $H \Delta A = \{X \Delta A \colon X \in H\}$. Show that

$$\mathrm{VCdim}(H \Delta A) = \mathrm{VCdim}(H).$$

3.18 Symmetric functions. A function $h \colon \{0,1\}^n \to \{0,1\}$ is *symmetric* if its value is uniquely determined by the number of 1's in the input. Let C denote the set of all symmetric functions.

(a) Determine the VC-dimension of C.

(b) Give lower and upper bounds on the sample complexity of any consistent PAC learning algorithm for C.

(c) Note that any hypothesis $h \in C$ can be represented by a vector $(y_0, y_1, ..., y_n) \in \{0,1\}^{n+1}$, where y_i is the value of h on examples having precisely i 1's. Devise a consistent learning algorithm for C based on this representation.

3.19 Biased coins. Professor Moent has two coins in his pocket, coin x_A and coin x_B. Both coins are slightly biased, i.e., $\Pr[x_A = 0] = 1/2 - \epsilon/2$ and $\Pr[x_B = 0] = 1/2 + \epsilon/2$, where $0 < \epsilon < 1$ is a small positive number, 0 denotes heads and 1 denotes tails. He likes to play the following game with his students. He picks a coin $x \in \{x_A, x_B\}$ from his pocket uniformly at random, tosses it m times, reveals the sequence of 0s and 1s he obtained and asks which coin was tossed. Determine how large m needs to be for a student's coin prediction error to be at most $\delta > 0$.

(a) Let S be a sample of size m. Professor Moent's best student, Oskar, plays according to the decision rule $f_o \colon \{0,1\}^m \to \{x_A, x_B\}$ defined by $f_o(S) = x_A$ iff $N(S) < m/2$, where $N(S)$ is the number of 0's in sample S.
Suppose m is even, then show that

$$error(f_o) \geq \frac{1}{2} \Pr\left[N(S) \geq \frac{m}{2} \middle| x = x_A\right]. \tag{3.56}$$

(b) Assuming m even, use the inequalities given in the appendix (section D.3)

to show that

$$error(f_o) > \frac{1}{4}\Big[1 - \Big[1 - e^{-\frac{m\epsilon^2}{1-\epsilon^2}}\Big]^{\frac{1}{2}}\Big]. \tag{3.57}$$

(c) Argue that if m is odd, the probability can be lower bounded by using $m+1$ in the bound in (a) and conclude that for both odd and even m,

$$error(f_o) > \frac{1}{4}\Big[1 - \Big[1 - e^{-\frac{2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}}\Big]^{\frac{1}{2}}\Big]. \tag{3.58}$$

(d) Using this bound, how large must m be if Oskar's error is at most δ, where $0 < \delta < 1/4$. What is the asymptotic behavior of this lower bound as a function of ϵ?

(e) Show that no decision rule $f\colon \{0,1\}^m \to \{x_a, x_B\}$ can do better than Oskar's rule f_o. Conclude that the lower bound of the previous question applies to all rules.

3.20 Infinite VC-dimension.

(a) Show that if a concept class C has infinite VC-dimension, then it is not PAC-learnable.

(b) In the standard PAC-learning scenario, the learning algorithm receives all examples first and then computes its hypothesis. Within that setting, PAC-learning of concept classes with infinite VC-dimension is not possible as seen in the previous question.
Imagine now a different scenario where the learning algorithm can alternate between drawing more examples and computation. The objective of this problem is to prove that PAC-learning can then be possible for some concept classes with infinite VC-dimension.
Consider for example the special case of the concept class C of all subsets of natural numbers. Professor Vitres has an idea for the first stage of a learning algorithm L PAC-learning C. In the first stage, L draws a sufficient number of points m such that the probability of drawing a point beyond the maximum value M observed be small with high confidence. Can you complete Professor Vitres' idea by describing the second stage of the algorithm so that it PAC-learns C? The description should be augmented with the proof that L can PAC-learn C.

3.21 VC-dimension generalization bound – realizable case. In this exercise we show that the bound given in corollary 3.4 can be improved to $O(\frac{d\log(m/d)}{m})$ in the realizable setting. Assume we are in the realizable scenario, i.e. the target concept is included in our hypothesis class H. We will show that if a hypothesis h is consistent

with a sample $S \sim D^m$ then for any $\epsilon > 0$ such that $m\epsilon \geq 8$

$$\Pr[R(h) > \epsilon] \leq 2\Big[\frac{2em}{d}\Big]^d 2^{-m\epsilon/2}\,. \tag{3.59}$$

(a) Let $H_S \subseteq H$ be the subset of hypotheses consistent with the sample S, let $\widehat{R}_S(h)$ denote the empirical error with respect to the sample S and define S' as a another independent sample drawn from D^m. Show that the following inequality holds for any $h_0 \in H_S$:

$$\Pr\Big[\sup_{h\in H_S} |\widehat{R}_S(h) - \widehat{R}_{S'}(h)| > \frac{\epsilon}{2}\Big] \geq \Pr\Big[B[m,\epsilon] > \frac{m\epsilon}{2}\Big]\Pr[R(h_0) > \epsilon]\,,$$

where $B[m,\epsilon]$ is a binomial random variable with parameters $[m,\epsilon]$. (*Hint*: prove and use the fact that $\Pr[\widehat{R}(h) \geq \frac{\epsilon}{2}] \geq \Pr[\widehat{R}(h) > \frac{\epsilon}{2} \wedge R(h) > \epsilon]$.)

(b) Prove that $\Pr\Big[B(m,\epsilon) > \frac{m\epsilon}{2}\Big] \geq \frac{1}{2}$. Use this inequality along with the result from (a) to show that for any $h_0 \in H_S$

$$\Pr\Big[R(h_0) > \epsilon\Big] \leq 2\Pr\Big[\sup_{h\in H_S} |\widehat{R}_S(h) - \widehat{R}_{S'}(h)| > \frac{\epsilon}{2}\Big]\,.$$

(c) Instead of drawing two samples, we can draw one sample T of size $2m$ then uniformly at random split it into S and S'. The right hand side of part (b) can then be rewritten as:

$$\Pr\Big[\sup_{h\in H_S} |\widehat{R}_S(h) - \widehat{R}_{S'}(h)| > \frac{\epsilon}{2}\Big] = \Pr_{\substack{T\sim D^{2m}:\\ T\to[S,S']}}\Big[\exists h \in H : \widehat{R}_S(h) = 0 \wedge \widehat{R}_{S'}(h) > \frac{\epsilon}{2}\Big]\,.$$

Let h_0 be a hypothesis such that $\widehat{R}_T(h_0) > \frac{\epsilon}{2}$ and let $l > \frac{m\epsilon}{2}$ be the total number of errors h_0 makes on T. Show that the probability of all l errors falling into S' is upper bounded by 2^{-l}.

(d) Part (b) implies that for any $h \in H$

$$\Pr_{\substack{T\sim D^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2} \ \Big|\ \widehat{R}_T(h_0) > \frac{\epsilon}{2}\Big] \leq 2^{-l}\,.$$

Use this bound to show that for any $h \in H$

$$\Pr_{\substack{T\sim D^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2}\Big] \leq 2^{-\frac{\epsilon m}{2}}\,.$$

(e) Complete the proof of inequality (3.59) by using the union bound to upper bound $\Pr_{\substack{T\sim D^{2m}:\\ T\to(S,S')}}\Big[\exists h \in H : \widehat{R}_S(h) = 0 \wedge \widehat{R}_{S'}(h) > \frac{\epsilon}{2}\Big]$. Show that we can achieve a high probability generalization bound that is of the order $O(\frac{d\log(m/d)}{m})$.

3.22 Generalization bound based on covering numbers. Let H be a family of functions mapping $\mathcal{X}$ to a subset of real numbers $\mathcal{Y} \subseteq \mathbb{R}$. For any $\epsilon > 0$, the *covering number* $\mathcal{N}(H, \epsilon)$ of H for the L_∞ norm is the minimal $k \in \mathbb{N}$ such that H can be covered with k balls of radius ϵ, that is, there exists $\{h_1, \ldots, h_k\} \subseteq H$ such that, for all $h \in H$, there exists $i \leq k$ with $\|h - h_i\|_\infty = \max_{x \in \mathcal{X}} |h(x) - h_i(x)| \leq \epsilon$. In particular, when H is a compact set, a finite covering can be extracted from a covering of H with balls of radius ϵ and thus $\mathcal{N}(H, \epsilon)$ is finite.

Covering numbers provide a measure of the complexity of a class of functions: the larger the covering number, the richer is the family of functions. The objective of this problem is to illustrate this by proving a learning bound in the case of the squared loss. Let D denote a distribution over $\mathcal{X} \times \mathcal{Y}$ according to which labeled examples are drawn. Then, the generalization error of $h \in H$ for the squared loss is defined by $R(h) = \mathrm{E}_{(x,y)\sim D}[(h(x) - y)^2]$ and its empirical error for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ by $\widehat{R}(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2$. We will assume that H is bounded, that is there exists $M > 0$ such that $|h(x) - y| \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The following is the generalization bound proven in this problem:

$$\Pr_{S\sim D^m}\left[\sup_{h\in H} |R(h) - \widehat{R}(h)| \geq \epsilon\right] \leq \mathcal{N}\Big(H, \frac{\epsilon}{8M}\Big) 2 \exp\Big(\frac{-m\epsilon^2}{2M^4}\Big). \tag{3.60}$$

The proof is based on the following steps.

(a) Let $L_S = R(h) - \widehat{R}(h)$, then show that for all $h_1, h_2 \in H$ and any labeled sample S, the following inequality holds:

$$|L_S(h_1) - L_S(h_2)| \leq 4M \|h_1 - h_2\|_\infty .$$

(b) Assume that H can be covered by k subsets $B_1, \ldots, B_k$, that is $H = B_1 \cup \ldots \cup B_k$. Then, show that, for any $\epsilon > 0$, the following upper bound holds:

$$\Pr_{S\sim D^m}\left[\sup_{h\in H} |L_S(h)| \geq \epsilon\right] \leq \sum_{i=1}^k \Pr_{S\sim D^m}\left[\sup_{h\in B_i} |L_S(h)| \geq \epsilon\right].$$

(c) Finally, let $k = \mathcal{N}(H, \frac{\epsilon}{8M})$ and let $B_1, \ldots, B_k$ be balls of radius $\epsilon/(8M)$ centered at $h_1, \ldots, h_k$ covering H. Use part (a) to show that for all $i \in [1, k]$,

$$\Pr_{S\sim D^m}\left[\sup_{h\in B_i} |L_S(h)| \geq \epsilon\right] \leq \Pr_{S\sim D^m}\left[|L_S(h_i)| \geq \frac{\epsilon}{2}\right],$$

and apply Hoeffding's inequality (theorem D.1) to prove (3.60).