CHAPTER FOURTEEN

Sample Complexity, VC Dimension, and Rademacher Complexity

Sampling is a powerful technique at the core of statistical data analysis and machine learning. Using a finite, often small, set of observations, we attempt to estimate properties of an entire sample space. How good are estimates obtained from a sample? Any rigorous application of sampling requires an understanding of the *sample complexity* of the problem – the minimum size sample needed to obtain the required results. In this chapter we focus on the sample complexity of two important applications of sampling: range detection and probability estimation. Here a range is just a subset of the underlying space. Our goal is to use one set of samples to detect a set of ranges or estimate the probabilities of a set of ranges, where the set of possible ranges is large, in fact possibly infinite. For detection, we mean that we want the sample to intersect with each range in the set, while for probability estimation, we want the fraction of points in the sample that intersect with each range in the set to approximate the probability associated with that range.

As an example, consider a sample $x_1, \ldots, x_m$ of m independent observations from an unknown distribution $\mathcal{D}$, where the values for our samples are in $\mathbb{R}$. Given an interval $[a, b]$, if the probability of the interval is at least ϵ, i.e., $\Pr(x \in [a, b]) \geq \epsilon$, then the probability that a sample of size $m = \frac{1}{\epsilon} \ln \frac{1}{\delta}$ intersects (or, in this context, detects) the interval $[a, b]$ is at least $1 - (1 - \epsilon)^m \geq 1 - \delta$. Given a set of k intervals, each of which has probability at least ϵ, we can apply a union bound to show that the probability that a sample of size $m' = \frac{1}{\epsilon} \ln \frac{k}{\delta}$ intersects each of the k intervals is at least $1 - k(1 - \epsilon)^{m'} \geq 1 - \delta$.

In many applications we need a sample that intersects with *every* interval that has probability at least ϵ, and there can be an infinite number of such intervals. What sample size guarantees that? We cannot use a simple union bound to answer this question, as our above analysis does not make sense when k is infinite. However, if there are many such intervals, there can be significant overlap between them. For example, consider samples chosen uniformly over $[0, 1]$ with $\epsilon = 1/10$; there are infinitely many intervals $[a, b]$ of length at least $1/10$, but the largest number of disjoint intervals of size at least $1/10$ is ten. A sample point may intersect with many intervals, and thus a small sample may be sufficient.

Indeed, the technique we will develop in this chapter will show that for any distribution $\mathcal{D}$, a sample of size $\Omega(\frac{1}{\epsilon} \ln \frac{1}{\delta})$, with probability at least $1 - \delta$, intersects all intervals of probability at least ϵ. Similarly, we will show that a sample of size $\Omega(\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$, with probability at least $1 - \delta$, simultaneously estimates the probabilities of all intervals, where each probability is estimated within an additive error bounded by ϵ.

The above example shows that the set of intervals on a line corresponds to a set of ranges that is easy to sample. In this chapter we develop general methods for evaluating the sample complexity of sets of ranges. We will see an example of sets of ranges with significantly larger sample complexity than the intervals example, and even sets of ranges with infinite sample complexity for either detection or probability estimation. We also present applications of the theory to rigorous machine learning and data mining analysis.

14.1. The Learning Setting

The study of sample complexity was motivated by statistical machine learning. To motivate our discussion of these concepts, we show how the task of learning a binary classification can be framed as either a detection or a probability estimation problem.

As a starting example, suppose that we know that a publisher uses a certain rule when determining whether to review or reject a book based on the submitted manuscript. The rule is a conjunction over certain Boolean variables (or their negations); for example, there could be a Boolean variable for whether the manuscript is over 100 pages, for whether the topic was of wide interest, for whether the author had suitable experience, and so on. As outsiders, we might not know the rule, and the question is whether we can learn the rule after seeing enough examples.

A second example involves learning the range of temperatures in which some electronic equipment is functioning correctly. We test the equipment at various temperatures: some are too low and some are too high, but in between there is an interval of temperatures in which the equipment is functioning correctly. The question is to determine an appropriate range of temperatures where the equipment functions.

Here is a general model for this sort of problem; we formalize these definitions later. We have a universe U of objects that we wish to classify, and let $c : U \to \{-1, 1\}$ be the correct, unknown classification. Usually $c(x) = 1$ corresponds to x being a "positive" example, and $c(x) = -1$ corresponds to x being a "negative" example. The correct classification also can be thought of as the subset of the universe corresponding to the positive examples.

The learning algorithm receives a training set $(x_1, c(x_1)), \ldots, (x_m, c(x_m))$, where $x_i \in U$ is chosen according to an unknown distribution $\mathcal{D}$, and $c(x_i)$ is the correct classification of x_i. The algorithm also receives a collection $\mathcal{C}$ of hypotheses, or possible classifications, to choose from. This collection of hypotheses can be referred to as the concept class. The output of the algorithm is a classification $h \in \mathcal{C}$. In the context of binary classification, every $h \in \mathcal{C}$ is also a function $h : U \to \{-1, 1\}$. Equivalently, each hypothesis is itself a subset of the universe, corresponding to the elements x with

$h(x) = 1$. The correctness of the chosen classification is evaluated with respect to its error in classifying new objects chosen according to the distribution $\mathcal{D}$.

In our first example, $\mathcal{C}$ is the collection of all possible conjunctions of subsets of the Boolean variables or their negations. That is, each $h \in \mathcal{C}$ corresponds to a Boolean formula given by a conjunction of variables; $h(x)$ is 1 if the Boolean expression evaluates to true on x, and -1 if it evaluates to false. In the second example, $\mathcal{C}$ is the set of all intervals in $\mathbb{R}$, so that for each $h \in \mathcal{C}$, $h(x) = 1$ if x is a point in the corresponding interval and $h(x) = -1$ otherwise.

Assume first that the correct classification c is included in the collection $\mathcal{C}$ of possible classifications. For any other $h \in \mathcal{C}$ let

$$\Delta(c, h) = \{x \in U \mid c(x) \neq h(x)\}$$

be the set of objects that are not classified correctly by classification h. The probability of a set $\Delta(c, h)$ is the probability that the distribution $\mathcal{D}$ generates an object in $\Delta(c, h)$. If our training set intersects with every set $\Delta(c, h)$ that has probability at least ϵ, then the learning algorithm can eliminate any classification $h \in \mathcal{C}$ that has error at least ϵ on input from $\mathcal{D}$. Thus, a sample (training set) that with probability $1 - \delta$ detects (or intersects with) all sets $\{\Delta(c, h) | \Pr_{\mathcal{D}}(\Delta(c, h)) \geq \epsilon, h \in C\}$ guarantees that such an algorithm outputs with probability $1 - \delta$ a classification that errs with probability bounded by ϵ.

A more realistic scenario is that no classification in $\mathcal{C}$ is perfectly correct. In that case, we require the algorithm to return a classification in $\mathcal{C}$ with an error probability that is no more than ϵ larger (with respect to $\mathcal{D}$) than any classification in $\mathcal{C}$. If our training set approximates all sets $\{\Delta(c, h) \mid h \in \mathcal{C}\}$ to within an additive error $\epsilon/2$, then the learning algorithm has sufficient information to eliminate any $h \in \mathcal{C}$ with error which is at least ϵ larger than the error of the best hypothesis in $\mathcal{C}$.

Finally, we note a major difference between the two examples above. Since the number of possible conjunctions over a bounded number of variables or their complements is bounded, the set of possible classifications in the first example is finite, and we can use standard techniques (union bound and Chernoff bound) to bound the size of the required sample (training set), though the bound may be loose. In the second example, the size of the concept class is not bounded and we need more advanced techniques to obtain a bound on the sample complexity. We present here two major techniques to evaluate the sample complexity, VC dimension and Rademacher complexity.

14.2. VC Dimension

We begin with the formal definitions, using the setting of intervals on a line to help explain them, and then consider other examples.

The Vapnik–Chervonenkis (VC) dimension is defined on range spaces.

Definition 14.1: *A* range space *is a pair* $(X, \mathcal{R})$ *where:*

1. *X is a (finite or infinite) set of* points*;*
2. *$\mathcal{R}$ is a family of subsets of X, called* ranges*.*

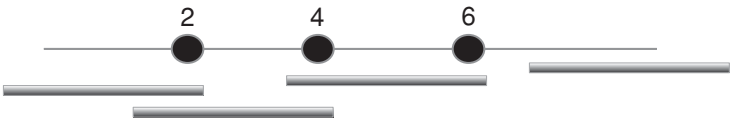

Figure 14.1: Let $\mathcal{R}$ be the collection of all closed intervals in $\mathbb{R}$. Any 2 points can be shattered, but there is no interval that separates $\{2, 6\}$ from $\{4\}$. The VC dimension of $(\mathbb{R}, \mathcal{R})$ is therefore 2.

If for example $X = \mathbb{R}$ is the set of real numbers, then $\mathcal{R}$ could be the family of all closed intervals $[a, b]$ in $\mathbb{R}$.

Given a set $S \subseteq X$, one can obtain a subset of S by intersecting it with a range $R \in \mathcal{R}$. The *projection* of $\mathcal{R}$ on S corresponds to the collection of all subsets that can be obtained in this way.

Definition 14.2: *Let $(X, \mathcal{R})$ be a range space and let $S \subseteq X$. The* projection *of $\mathcal{R}$ on S is*

$$\mathcal{R}_S = \{R \cap S \mid R \in \mathcal{R}\}.$$

For example, let $X = \mathbb{R}$ and $\mathcal{R}$ be the set of all closed intervals. Consider $S = \{2, 4\}$. The intersection of S with the interval $[0, 1]$ gives the empty set; the intersection of S with the interval $[1, 3]$ is $\{2\}$; the intersection of S with the interval $[3, 5]$ is $\{4\}$; and the intersection of S with the interval $[1, 5]$ is $\{2, 4\}$. Hence the projection of $\mathcal{R}$ on S is the set of all possible subsets of S in this case, and indeed the same is true for any set of two distinct points.

Consider now a set $S = \{2, 4, 6\}$. You should convince yourself that the projection of $\mathcal{R}$ on S includes seven of the eight subsets of S, but not $\{2, 6\}$. This is because an interval containing 2 and 6 must also contain 4. More generally, the projection of $\mathcal{R}$ on any set S of three distinct points would contain only seven of the eight possible subsets of S.

We measure the complexity of a range space $(X, \mathcal{R})$ by considering the largest subset S of X such that all subsets of S are contained in the projection of $\mathcal{R}$ on S.

Definition 14.3: *Let $(X, \mathcal{R})$ be a range space. A set $S \subseteq X$ is* shattered *by $\mathcal{R}$ if $|\mathcal{R}_S| = 2^{|S|}$. The* Vapnik–Chervonenkis (VC) dimension *of a range space $(X, \mathcal{R})$ is the maximum cardinality of a set $S \subseteq X$ that is shattered by $\mathcal{R}$. If there are arbitrarily large finite sets that are shattered by $\mathcal{R}$, then the VC dimension is infinite.*

We have shown that any set of two points is shattered by closed intervals on the real number line, but that any set of three points is not. Of course, that argument also shows that no larger set of points is shattered by closed intervals. Therefore, the VC dimension of that range space is 2. Our example shows that a range space with an infinite set of points and an infinite number of ranges can have a bounded VC dimension. (See Figure 14.1.)

An important subtlety in the definition is that the VC dimension of a range space is d if there is *some* set of cardinality d that is shattered by $\mathcal{R}$. It does not imply that all sets of cardinality d are shattered by $\mathcal{R}$. On the other hand, to show that the VC dimension

Figure 14.2: Let $\mathcal{R}$ be the collection of all half-space partitions on $\mathbb{R}^2$. Any three points can be shattered, but there is no half-space partition that separates the two white points from the two black points. Thus, the VC dimension of $(\mathbb{R}^2, \mathcal{R})$ is 3.

is not $d + 1$ or larger, one must show that *all* sets of cardinality larger than d are not shattered by $\mathcal{R}$.

14.2.1. *Additional Examples of VC Dimension*

We consider some other simple examples of VC dimension.

Linear half-spaces
Let $X = \mathbb{R}^2$ and let $\mathcal{R}$ be the set of all half-spaces defined by a linear partition of the plane. That is, we consider all possible lines $ax + by = c$ in the plane, and $\mathcal{R}$ consists of all half-spaces $ax + by \geq c$. The VC dimenstion in this case is at least 3, since any set of three points that do not lie on a line can be shattered. On the other hand, no set of four points can be shattered. To see this, we need to consider several cases. First, if any three points lie on a line they cannot be shattered, as we cannot separate the middle point from the other two by any half-space. Hence we may assume no three points lie on a line; this is often referred to as the points being in "general position". Second, if one point lies within the convex hull defined by the other three points, no half-space can separate that point from the other three. Finally, if the four points define a convex hull, then there is no half-space that separates two non-neighboring points from the other two. (See Figure 14.2.)

While harder to visualize, if $X = \mathbb{R}^d$ and $\mathcal{R}$ corresponds to all half-spaces in d dimensions, the VC dimension is $d + 1$. (See Exercise 14.7.)

Convex sets
Let $X = \mathbb{R}^2$ and let $\mathcal{R}$ be the family of all closed convex sets on the plane. We show that this range space has infinite VC dimension by showing that for every n there exists a set of size n that can be shattered. Let $S_n = \{x_1, \ldots, x_n\}$ be a set of n points on the boundary of a circle. Any subset $Y \subseteq S_n, Y \neq \emptyset$ defines a convex set that does not include any point in $S_n \setminus Y$, and hence Y is included in the projection of $\mathcal{R}$ on S_n. The empty set is easily seen to be in the projection as well. Hence, for any number of points n, the set S_n is shattered and the VC dimension is therefore infinite. (See Figure 14.3.)

Monotone Boolean conjunctions
Let $y_1, y_2, \ldots, y_n$ be n Boolean variables, and let MC_n be the collection of functions defined by conjunctions of subsets of the non-negated variables y_i. Let $X = \{0, 1\}^n$

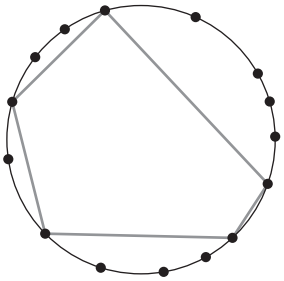

Figure 14.3: Let $\mathcal{R}$ be the set of all convex bodies in $\mathbb{R}^2$. Any partition of the set of points on the circle can be defined by a convex body. Therefore, the VC dimension of $(\mathbb{R}^2, \mathcal{R})$ is infinite.

correspond to all possible truth assignments of the n variables in the natural way. For each function $f \in MC_n$ let $R_f = \{\bar{a} \in X \ : \ f(\bar{a}) = 1\}$ be the set of inputs that satisfy f, and let $\mathcal{R} = \{R_f \mid f \in MC_n\}$. Consider the set $S \subseteq X$ of n points:

$$
\begin{gathered}
(0, 1, 1, \ldots, 1) \\
(1, 0, 1, \ldots, 1) \\
(1, 1, 0, \ldots, 1) \\
\vdots \\
(1, 1, 1, \ldots, 0).
\end{gathered}
$$

We claim that each subset of S is equal to $S \cap R_f$ for some R_f. For example, the complete set S corresponds to $S \cap R_f$ for the trivial function that is always 1, i.e., $f(\bar{a}) = 1$. More generally, the subset of S that has all points except those with a 0 in coordinates $i_1, i_2, \ldots, i_j$ is equal to $S \cap R_f$ for $f(\bar{a}) = y_{i_1} \wedge y_{i_2} \wedge \cdots \wedge y_{i_j}$. This set can therefore be shattered by $\mathcal{R}$ and the VC dimension is at least n. The VC dimension cannot be larger than n since $|\mathcal{R}| = |MC_n| = 2^n$, and hence there can be at most 2^n distinct intersections of the form $S \cap R_f$. If the VC dimension was larger than n, at least 2^{n+1} different intersections would be needed.

14.2.2. *Growth Function*

The combinatorial significance of the concept of the VC dimension is that it gives a bound on the number of different ranges in the projection of the range space on a smaller set of points. In particular, when a range space with finite VC dimension $d \geq 2$ is projected on a set of n points, the number of different ranges in the projection is bounded by a polynomial in n with maximum degree d.

To prove this property we define the *growth function*

$$
\mathcal{G}(d, n) = \sum_{i=0}^{d} \binom{n}{i}.
$$

For $n = d$, we have $\mathcal{G}(d, n) = 2^d$, and for $n > d \geq 2$, we have

$$\mathcal{G}(d, n) \leq \sum_{i=0}^{d} \frac{n^i}{i!} \leq n^d.$$

The growth function is related to the VC dimension through the following theorem.

Theorem 14.1 [Sauer–Shelah]: *Let $(X, \mathcal{R})$ be a range space with $|X| = n$ and VC dimension d. Then $|\mathcal{R}| \leq \mathcal{G}(d, n)$.*

Proof: We prove the claim by induction on d, and for each d by induction on n. As the base case, the claim clearly holds for $d = 0$ or $n = 0$, as in both of these cases $G(d, n) = 1$, with the only possible $\mathcal{R}$ being the family containing only the empty set.

Assume that the claim holds for $d - 1$ and $n - 1$, and for d and $n - 1$. We may therefore assume $|X| = n > 0$. For some $x \in X$, consider two range spaces on $X \setminus \{x\}$:

$$\mathcal{R}_1 = \{R \setminus \{x\} \mid R \in \mathcal{R}\}$$

and

$$\mathcal{R}_2 = \{R \setminus \{x\} \mid R \cup \{x\} \in \mathcal{R} \text{ and } R \setminus \{x\} \in \mathcal{R}\}.$$

We first observe that $|\mathcal{R}| = |\mathcal{R}_1| + |\mathcal{R}_2|$. Indeed, each set $R \in \mathcal{R}$ is mapped to a set $R \setminus \{x\} \in \mathcal{R}_1$, but if both $R \cup \{x\}$ and $R \setminus \{x\}$ are in $\mathcal{R}$, then both sets are mapped to the same set $R \setminus \{x\} \in \mathcal{R}_1$. By including that set again in $\mathcal{R}_2$, we have $|\mathcal{R}| = |\mathcal{R}_1| + |\mathcal{R}_2|$.

Now $(X \setminus \{x\}, \mathcal{R}_1)$ is a range space on $n - 1$ items, and its VC dimension is bounded above by d, the VC dimension of $(X, \mathcal{R})$. To see this, assume that $\mathcal{R}_1$ shatters a set S of size $d + 1$ in $X \setminus \{x\}$. Then S is also shattered by $\mathcal{R}$, as for any $R \in \mathcal{R}_1$, there is a corresponding R' in $\mathcal{R}$ that is either R or $R \cup \{x\}$, and in either case the projection of $\mathcal{R}$ on S contains $S \cap R' = S \cap R$. But then $\mathcal{R}$ would shatter the set S, contradicting the assumption that $(X, \mathcal{R})$ has VC dimension d.

Similarly, $(X \setminus \{x\}, \mathcal{R}_2)$ is a range space on $n - 1$ items, and its VC dimension is bounded above by $d - 1$. To see this, assume that $\mathcal{R}_2$ shatters a set S of size d in $X \setminus \{x\}$. Then consider the set $S \cup \{x\}$ in $\mathcal{R}$. For any $R \in \mathcal{R}_2$, both R and $R \cup \{x\}$ are in $\mathcal{R}$, and hence one can obtain both $(S \cup \{x\}) \cap R = S \cap R$ and $(S \cup \{x\}) \cap (R \cup \{x\}) = S \cup \{x\}$ in the projection of $\mathcal{R}$ on S. But then $\mathcal{R}$ would shatter the set $S \cup \{x\}$, contradicting the assumption that $(X, \mathcal{R})$ has VC dimension d.

Applying the induction hypothesis we get

$$\begin{aligned}
|\mathcal{R}| = |\mathcal{R}_1| + |\mathcal{R}_2| &\leq \mathcal{G}(d, n-1) + \mathcal{G}(d-1, n-1) \\
&\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= 1 + \sum_{i=0}^{d-1} \left(\binom{n-1}{i+1} + \binom{n-1}{i} \right) \\
&= \sum_{i=0}^{d} \binom{n}{i} = \mathcal{G}(d, n).
\end{aligned}$$

■

14.2.3. *VC dimension component bounds*

We can sometimes bound the VC dimension of a complex range space as a function of the VC dimension of its simpler components.

The projection of a range space $(X, \mathcal{R})$ on a set $Y \subseteq X$ defines a range space $(Y, \mathcal{R}_Y)$ with $\mathcal{R}_Y = \{R \cap Y \mid R \in \mathcal{R}\}$. We have the following corollary of Theorem 14.1.

Corollary 14.2: *Let $(X, \mathcal{R})$ be a range space with VC dimension d, and let $Y \subseteq X$. Then*

$$|\mathcal{R}_Y| \leq \mathcal{G}(d, |Y|).$$

We also require the following technical lemma.

Lemma 14.3: *If $y \geq x \ln x \geq \mathrm{e}$, then $\frac{2y}{\ln y} \geq x$.*

Proof: For $y = x \ln x$ we have $\ln y = \ln x + \ln \ln x \leq 2 \ln x$. Thus

$$\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.$$

Differentiating $f(y) = \frac{\ln y}{2y}$ we find that $f(y)$ is monotonically decreasing when $y \geq x \ln x \geq \mathrm{e}$, and hence $\frac{2y}{\ln y}$ is monotonically increasing on the same interval, proving the lemma. ■

We are now ready for the following theorem.

Theorem 14.4: *Let $(X, \mathcal{R}^1), \ldots, (X, \mathcal{R}^k)$ be k range spaces, each with VC dimension at most d. Let $f : (\mathcal{R}^1, \ldots, \mathcal{R}^k) \to 2^X$ be a mapping of k-tuples $(r_1, \ldots, r_k) \in (\mathcal{R}^1, \ldots, \mathcal{R}^k)$ to subsets of X, and let*

$$\mathcal{R}^f = \{f(r_1, \ldots, r_k) \mid r_1 \in \mathcal{R}^1, \ldots, r_k \in \mathcal{R}^k\}.$$

The VC dimension of the range space $(X, \mathcal{R}^f)$ is $O(kd \ln(kd))$.

Proof: Let the VC dimension of $(X, \mathcal{R}^f)$ be at least t, so there is a set $Y \subseteq X$ shattered by $\mathcal{R}^f$ with $t = |Y|$. Since the VC dimension of $(X, \mathcal{R}^i)$, $1 \leq i \leq k$, is at most d, by Corollary 14.2, $|\mathcal{R}^i_Y| \leq \mathcal{G}(d, t) \leq t^d$. Thus, the number of subsets in the projection of $\mathcal{R}^f$ on Y is bounded by

$$|\mathcal{R}^f_Y| \leq |\mathcal{R}^1_Y| \times \cdots \times |\mathcal{R}^k_Y| \leq t^{dk}.$$

Since $\mathcal{R}^f_Y$ shatters Y, $|\mathcal{R}^f_Y| \geq 2^t$. Hence $t^{dk} \geq 2^t$. Let us assume that $y \geq x \ln x$ for $y = t$ and $x = \frac{2(dk+1)}{\ln 2}$ and derive a contradiction. Applying Lemma 14.3,

$$\frac{2y}{\ln y} = \frac{2t}{\ln t} \geq \frac{2(dk+1)}{\ln 2}.$$

It follows that

$$t \geq (dk+1) \log_2 t,$$

so $2^t \geq t^{dk+1} > t^{kd}$. Hence if $t \geq x \ln x$, which is $\Omega(dk \ln(dk))$, we have a contradiction. It follows that t must be $O(kd \ln(kd))$. ■

The following stronger result below is proven in Exercise 14.10.

Theorem 14.5: *Let $(X, \mathcal{R}^1), \ldots, (X, \mathcal{R}^k)$ be k range spaces each with VC dimensions at most d. Let $f : (\mathcal{R}^1, \ldots, \mathcal{R}^k) \to 2^X$ be a mapping of k-tuples $(r_1, \ldots, r_k) \in (\mathcal{R}^1, \ldots, \mathcal{R}^k)$ to subsets of X, and let*

$$\mathcal{R}^f = \{f(r_1, \ldots, r_k) \mid r_1 \in \mathcal{R}^1, \ldots, r_k \in \mathcal{R}^k\}.$$

The VC dimension of the range space $(X, \mathcal{R}^f)$ is $O(kd \ln k)$.

This yields the following corollary.

Corollary 14.6: *Let $(X, \mathcal{R}^1)$ and $(X, \mathcal{R}^2)$ be two range spaces, each with VC dimension at most d. Let*

$$\mathcal{R}^{\cup} = \{r_1 \cup r_2 \mid r_1 \in \mathcal{R}^1 \text{ and } r_2 \in \mathcal{R}^2\},$$

and

$$\mathcal{R}^{\cap} = \{r_1 \cap r_2 \mid r_1 \in \mathcal{R}^1 \text{ and } r_2 \in \mathcal{R}^2\}.$$

The VC dimensions of the range spaces $(X, \mathcal{R}^{\cup})$ and $(X, \mathcal{R}^{\cap})$ are $O(d)$.

14.2.4. *ϵ-nets and ϵ-samples*

The applications of VC dimension to sampling, including to the types of learning problems mentioned at the beginning of the chapter, can be formulated in terms of objects called ϵ-nets and ϵ-samples.

As a combinatorial object, an ϵ-net for a subset $A \subseteq X$ of a range space is a subset $N \subseteq A$ of points that intersects with all ranges in the range space that are not too small with respect to A, in that the range contains an ϵ-fraction of A. The object is called a net because it "catches," or intersects, every range of sufficient size.

Definition 14.4 [combinatorial definition]: *Let $(X, \mathcal{R})$ be a range space, and let $A \subseteq X$ be a finite subset of X. A set $N \subseteq A$ is a* combinatorial ϵ-net *for A if N has a nonempty intersection with every set $R \in \mathcal{R}$ such that $|R \cap A| \geq \epsilon |A|$.*

However, ϵ-nets can also be defined more generally with respect to a distribution $\mathcal{D}$ on the point set X. The combinatorial definition above corresponds to a setting where the distribution $\mathcal{D}$ is uniform over the set A. The more general form below is more useful for many algorithmic applications. In what follows, recall that $\Pr_{\mathcal{D}}(R)$ for a set R is the probability that a point chosen according to $\mathcal{D}$ is in R.

Definition 14.5: *Let $(X, \mathcal{R})$ be a range space, and let $\mathcal{D}$ be a probability distribution on X. A set $N \subseteq X$ is an* ϵ-net *for X with respect to $\mathcal{D}$ if for any set $R \in \mathcal{R}$ such that $\Pr_{\mathcal{D}}(R) \geq \epsilon$, the set R contains at least one point from N, i.e.,*

$$\forall R \in \mathcal{R}, \ \Pr_{\mathcal{D}}(R) \geq \epsilon \Rightarrow R \cap N \neq \emptyset.$$

An ϵ-sample (also called an ϵ-approximation) provides even stronger guarantees than an ϵ-net. It not only intersects every suitably large range, but also ensures that every range has roughly the right relative frequency within the sample.

Definition 14.6: *Let $(X, \mathcal{R})$ be a range space, and let $\mathcal{D}$ be a probability distribution on X. A set $S \subseteq X$ is an* ϵ-sample *for X with respect to $\mathcal{D}$ if for all sets $R \in \mathcal{R}$,*

$$\left| \Pr_{\mathcal{D}}(R) - \frac{|S \cap R|}{|S|} \right| \leq \epsilon.$$

Again, by fixing the distribution $\mathcal{D}$ to be uniform over a finite set $A \subseteq X$, we obtain the combinatorial version of this concept.

Definition 14.7 [combinatorial definition]: *Let $(X, \mathcal{R})$ be a range space, and let $A \subseteq X$ be a finite subset of X. A set $N \subseteq A$ is a* combinatorial ϵ-sample *for A if for all sets $R \in \mathcal{R}$,*

$$\left| \frac{|A \cap R|}{|A|} - \frac{|N \cap R|}{|N|} \right| \leq \epsilon.$$

In what follows, we may say ϵ-net and ϵ-sample in place of the more exact terms combinatorial ϵ-net and combinatorial ϵ-sample when the meaning should be clear from context.

Our goal is to obtain ϵ-nets and ϵ-samples through sampling. We say that a set S is a sample of size m from a distribution $\mathcal{D}$ if the m elements of S were chosen independently with distribution $\mathcal{D}$.

Definition 14.8: *A range space $(X, \mathcal{R})$ has the* uniform convergence property *if for every $\epsilon, \delta > 0$ there is a sample size $m = m(\epsilon, \delta)$ such that for every distribution $\mathcal{D}$ over X, if S is a random sample from $\mathcal{D}$ of size m then, with probability at least $1 - \delta$, S is an ϵ-sample for X with respect to $\mathcal{D}$.*

In the following sections we show that the minimum sample size that contains an ϵ-net or an ϵ-sample for a range space can be bounded in terms of the VC dimension of the range space, independent of the numbers of its points or ranges. In particular, we will show that a range space has the uniform convergence property if and only if its VC dimension is finite. These results show that the VC dimension is a concrete, useful measure of the complexity of a range space.

14.3. The ϵ-net Theorem

As a first step, we use a standard union bound argument to obtain bounds on the size of a combinatorial ϵ-net via the probabilistic method.

Theorem 14.7: *Let $(X, \mathcal{R})$ be a range space with VC dimension $d \geq 2$ and let $A \subseteq X$ have size $|A| = n$. Then there exists a combinatorial ϵ-net N for A of size at most $\lceil \frac{d \ln n}{\epsilon} \rceil$.*

Proof: Consider the projection of the range space $\mathcal{R}$ on A; denote this by $\mathcal{R}'$. By Theorem 14.1, the size of $\mathcal{R}'$ is at most $\mathcal{G}(d, n) \leq n^d$.

Suppose we take a sample of $k = \lceil \frac{d \ln n}{\epsilon} \rceil$ points of A independently and uniformly at random. For each set $R \in \mathcal{R}$ such that $|R \cap A| \geq \epsilon |A|$, there is a corresponding set $R' \in \mathcal{R}'$. The probability that our sample misses a given set R' is $(1 - \epsilon)^k$, and there are

at most n^d possible sets R' to consider. Applying a union bound, the probability that the sample misses at least one such R' is at most

$$n^d(1-\epsilon)^k < n^d \mathrm{e}^{-d\ln n} = 1.$$

Since the probability that a random sample of size $k = \lceil \frac{d\ln n}{\epsilon} \rceil$ misses at least one set R' is strictly less than 1, by the probabilistic method there is a set of that size that misses no set $R' \in \mathcal{R}'$, and is therefore an ϵ-net for A. ■

We can, however, in general do much better than the bound of Theorem 14.7. Our goal is to show that with high probability we can obtain an ϵ-net from a random sample of elements where the size of the sample does not depend on n, *as long as the VC dimension is finite*. This may appear somewhat surprising; while $O(1/\epsilon)$ points on average are needed to hit any particular range, it is not clear how to hit all of them without some dependence on n. Essentially, we are finding that the union bound of Theorem 14.7 is too weak an approach in this setting, and that the VC dimension provides a means to avoid it.

The following theorem, whose proof takes a somewhat unusual path that we sometimes refer to as "double sampling", provides our main results on ϵ-nets. The theorem holds for our more general notion of ϵ-nets, not just combinatorial ϵ-nets.

Theorem 14.8: *Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \delta, \epsilon \leq 1/2$, there is an*

$$m = O\left(\frac{d}{\epsilon}\ln\frac{d}{\epsilon} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-net for X with probability at least $1-\delta$.

In particular, Theorem 14.8 implies that there exists an ϵ-net of size $O(\frac{d}{\epsilon}\ln\frac{d}{\epsilon})$.

Proof: Let M be a set of m independent samples from X according to $\mathcal{D}$, and let E_1 be the event that M is not an ϵ-net for X with respect to the distribution $\mathcal{D}$, i.e.,

$$E_1 = \{\exists R \in \mathcal{R} \mid \Pr_{\mathcal{D}}(R) \geq \epsilon \text{ and } |R \cap M| = 0\}.$$

We want to show that $\Pr(E_1) \leq \delta$ for a suitable m. Notice that for any particular R, since $\Pr_{\mathcal{D}}(R) \geq \epsilon$, the expected size of $|R \cap M|$ would be at least ϵm, and hence it seems natural that $\Pr(E_1)$ is small. However, as the union bound argument of Theorem 14.7 is too weak to provide this strong a bound, we use an indirect means to bound $\Pr(E_1)$.

To do this, we choose a second set T of m independent samples from X according to $\mathcal{D}$ and define E_2 to be the event that some range R with $\Pr_{\mathcal{D}}(R) \geq \epsilon$ has an empty intersection with M but a reasonably large intersection with T:

$$E_2 = \{\exists R \in \mathcal{R} \mid \Pr_{\mathcal{D}}(R) \geq \epsilon \text{ and } |R \cap M| = 0 \text{ and } |R \cap T| \geq \epsilon m/2\}.$$

Since T is a random sample and $\Pr_{\mathcal{D}}(R) \geq \epsilon$, the event $|R \cap T| \geq \epsilon m/2$ should occur with nontrivial probability and therefore the events E_1 and E_2 should have similar probability. The following lemma formalizes this intuition:

Lemma 14.9: *For $m \geq 8/\epsilon$,*

$$\Pr(E_2) \leq \Pr(E_1) \leq 2\Pr(E_2).$$

Proof: As the event E_2 is included in the event E_1, we have $\Pr(E_2) \leq \Pr(E_1)$. For the second inequality, note that if event E_1 holds, there is some particular R' so that $|R' \cap M| = 0$ and $\Pr_{\mathcal{D}}(R') \geq \epsilon$. We use the definition of conditional probability to obtain

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr(|T \cap R'| \geq \epsilon m/2).$$

Now for a fixed range R' and a random sample T the random variable $|T \cap R'|$ has a binomial distribution $B(m, \Pr_{\mathcal{D}}(R'))$. Since $\Pr_{\mathcal{D}}(R') \geq \epsilon$, by applying the Chernoff bound (Theorem 4.5), we have for $m \geq 8/\epsilon$,

$$\Pr(|T \cap R'| < \epsilon m/2) \leq \mathrm{e}^{-\epsilon m/8} < 1/2.$$

Thus,

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr(|T \cap R'| \geq \epsilon m/2) \geq 1/2,$$

giving $\Pr(E_1) \leq 2\Pr(E_2)$ as desired. ■

The lemma above gives us an approach to showing that $\Pr(E_1)$ is small. The intuition is as follows: since M and T are both random samples of size m, it would be very surprising to have $|M \cap R| = 0$ but $|T \cap R|$ be large for some R. If we think of first sampling the m items that form M and then sampling the m items that form T, we must have somehow been very unlucky to have all the samples that intersect R come in the second set of m samples, and none in the first.

Formally, we bound the probability of E_2 by the probability of a larger event E_2':

$$E_2' = \{\exists R \in \mathcal{R} \mid |R \cap M| = 0 \text{ and } |R \cap T| \geq \epsilon m/2\}.$$

The event E_2' excludes the condition that $\Pr_{\mathcal{D}}(R) \geq \epsilon$; in some sense, that has been replaced by the condition on the size of $|R \cap T|$. The event E_2' now depends only on the elements in $M \cup T$.

Lemma 14.10: *It holds that*

$$\Pr(E_1) \leq 2\Pr(E_2) \leq 2\Pr(E_2') \leq 2(2m)^d 2^{-\epsilon m/2}.$$

Proof: Since M and T are random samples, we can assume that we first choose a set of $2m$ elements and then partition it randomly into two equal size sets M and T.

For a fixed $R \in \mathcal{R}$ and $k = \epsilon m/2$, let

$$E_R = \{|R \cap M| = 0 \text{ and } |R \cap T| \geq k\}.$$

To bound the probability of E_R we note that this event implies that $M \cup T$ has at least k elements of R, but all these elements were placed in T by the random partition. That is,

of the $\binom{2m}{m}$ possible partitions of $M \cup T$, we chose one of the $\binom{2m-k}{m}$ partitions where no element of R is in M.

Hence

$$\begin{aligned}\Pr(E_R) &\leq \Pr(|M \cap R| = 0 \mid |R \cap (M \cup T)| \geq k) \\ &= \frac{\binom{2m-k}{m}}{\binom{2m}{m}} \\ &= \frac{(2m-k)!m!}{(2m)!(m-k)!} \\ &= \frac{m(m-1)\cdots(m-k+1)}{(2m)(2m-1)\cdots(2m-k+1)} \\ &\leq 2^{-\epsilon m/2}.\end{aligned}$$

Our bound on $\Pr(E_R)$ does not depend on the choice of the set $T \cup M$, only on its random partition into T and M. By Theorem 14.1 the projection of $\mathcal{R}$ on $M \cup T$ has no more than $(2m)^d$ ranges. Thus,

$$\Pr(E_2') \leq (2m)^d 2^{-\epsilon m/2}. \qquad \blacksquare$$

To complete the proof of Theorem 14.8 we show that for

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{2}{\delta},$$

we have

$$\Pr(E_1) \leq 2 \Pr(E_2') \leq 2(2m)^d 2^{-\epsilon m/2} \leq \delta.$$

Equivalently, we require

$$\epsilon m/2 \geq \ln(2/\delta) + d \ln(2m).$$

Clearly it holds that $\epsilon m/4 \geq \ln(2/\delta)$, since $m > \frac{4}{\epsilon} \ln \frac{2}{\delta}$. It therefore suffices to show that $\epsilon m/4 \geq d \ln(2m)$ to complete the proof.

Applying Lemma 14.3 with $y = 2m \geq \frac{16d}{\epsilon} \ln \frac{16d}{\epsilon}$ and $x = \frac{16d}{\epsilon}$, we have

$$\frac{4m}{\ln(2m)} \geq \frac{16d}{\epsilon},$$

so

$$\frac{\epsilon m}{4} \geq d \ln(2m)$$

as required. $\blacksquare$

The above theorem gives a near tight bound, as shown by the following theorem (see Exercise 14.13 for a proof).

Theorem 14.11: *A random sample of a range space with VC dimension d that, with probability at least $1 - \delta$, is an ϵ-net must have size $\Omega(\frac{d}{\epsilon})$.*

14.4. Application: PAC Learning

Probably Approximately Correct (PAC) Learning provides a framework for mathematical analysis of computational learning from examples. PAC characterizes the complexity of a learning problem in terms of the number of examples and computation needed to provide answers that are approximately correct, in that they are approximately correct with good probability, on as yet unseen examples. We use the model of PAC learning to demonstrate an application of VC dimension to learning theory. However, we note that the VC dimension technique applies to a broader setting of statistical machine learning.

We turn now to a formal definition of PAC learning. We assume a set of items X and a probability distribution $\mathcal{D}$ defined on X. We work here in the setting of binary classifications, where a *concept* (or *classification*) can be treated as a subset $C \subseteq X$; all items in C are said to have a positive classification and all items in $X \setminus C$ are said to have a negative classification. Equivalently, a classification can be treated as a function $c(x)$ that is 1 if $x \in C$ and -1 if $x \notin C$. We use both notions of a classification interchangeably, where the meaning is clear. The *concept class* $\mathcal{C}$ is the set of all possible classifications defined by the problem.

The learning algorithm calls a function ORACLE that produces a pair $(x, c(x))$, where x is distributed according to $\mathcal{D}$, and $c(x)$ is 1 if $x \in C$ and -1 otherwise. We assume that successive calls to ORACLE are independent. For clarity, we may write ORACLE$(C, \mathcal{D})$ to specify the concept and distribution under consideration. We also assume that the classification problem is *realizable*, i.e. there is a classification $h \in \mathcal{C}$ that conforms with our input distribution. Formally,

$$\exists\, h \in \mathcal{C} \text{ such that } \Pr_{\mathcal{D}}(h(x) \neq c(x)) = 0.$$

We now define what it means for a concept to be learnable.

Definition 14.9 [PAC Learning]: *A concept class $\mathcal{C}$ over input set X is* PAC learnable[1] *if there is an algorithm L, with access to a function* ORACLE$(C, \mathcal{D})$*, that satisfies the following properties: for every correct concept $C \in \mathcal{C}$, every distribution $\mathcal{D}$ on X, and every $0 < \epsilon, \delta \leq 1/2$, the number of calls that the algorithm L makes to the function* ORACLE$(C, \mathcal{D})$ *is polynomial in ϵ^{-1} and δ^{-1}, and with probability at least $1 - \delta$ the algorithm L outputs a hypothesis h such that* $\Pr_{\mathcal{D}}(h(x) \neq c(x)) \leq \epsilon$.

We first prove that any finite concept class is PAC learnable.

Theorem 14.12: *Any finite concept class $\mathcal{C}$ can be PAC learned with $m = \frac{1}{\epsilon}(\ln |\mathcal{C}| + \ln \frac{1}{\delta})$ samples.*

Proof: Let $c^* \in \mathcal{C}$ be the correct classification. A hypothesis h is said to be "bad" if $\Pr_{\mathcal{D}}(h(x) \neq c^*(x)) \geq \epsilon$. The probability that any particular bad hypothesis is consistent

[1] PAC learning is mainly concerned with the computational complexity of learning. In particular, a concept class $\mathcal{C}$ is *efficiently PAC learnable* if the algorithm runs in time polynomial in the size of the problem, $1/\epsilon$ and $1/\delta$. Such an algorithm uses at most polynomially many samples. Here we are only interested in the sample complexity of the learning process; however, we note that the computational complexity of the learning algorithm is not necessarily polynomial in the sample size.

with m random samples is bounded above by $(1-\epsilon)^m$, and hence the probability that any bad hypothesis is consistent with m random samples is bounded above by

$$|\mathcal{C}|(1-\epsilon)^m \leq \delta.$$

The result follows. ■

We can also apply the PAC learning framework to infinite concept classes. Let us consider learning an interval $[a, b] \in \mathbb{R}$. The concept class here is the collection of all closed intervals in $\mathbb{R}$:

$$\mathcal{C} = \{[x, y] \mid x \leq y\} \cup \emptyset.$$

Notice that we also include a trivial concept that corresponds to the empty interval.

Let $c^* \in \mathcal{C}$ be the concept to be learned, and h be the hypothesis returned by our algorithm. The training set is a collection of n points drawn from a distribution $\mathcal{D}$ on $\mathbb{R}$, where each point in the interval $[a, b]$ is a positive example and each point outside the interval is a negative example. If none of the sample points are positive examples, then our algorithm returns the trivial hypothesis, where $h(x) = -1$ everywhere. If any of the sample points are positive examples, then let c and d respectively be the smallest and largest values of positive examples. Our algorithm then returns the interval $[c, d]$ as its hypothesis. (If there is only one positive example, the algorithm will return an interval of the form $[c, c]$.) By design, our algorithm can only make an error on an input x if $x \in [a, b]$; our algorithm will not make an error outside this interval, because it always returns -1 for points $x \notin [a, b]$.

We now determine the probability that our algorithm returns a bad hypothesis. Let us first consider the case where $\Pr_{\mathcal{D}}(x \in [a, b]) \leq \epsilon$. Because our algorithm can only return an incorrect answer on points in the interval $[a, b]$, our algorithm always returns a hypothesis with a probability of error at most ϵ in this case, and hence never returns a bad hypothesis.

Now let us consider when $\Pr_{\mathcal{D}}(x \in [a, b]) > \epsilon$. In this case, let $a' \geq a$ be the smallest value such that $\Pr_{\mathcal{D}}([a, a']) \geq \epsilon/2$. Similarly, let $b' \leq b$ be the largest value such that $\Pr_{\mathcal{D}}([b', b]) \geq \epsilon/2$. Here $a' \leq b'$ since $\Pr_{\mathcal{D}}(x \in [a, b]) > \epsilon$. For convenience, we assume $a' < b'$; the case $a' = b'$ can be handled similarly. (If $a' = b'$, then the point a' has nonzero probability of being selected, and we can divide up that probability among the intervals $[a, a']$ and $[b', b]$ so the probability of each is at least $\epsilon/2$.) For our algorithm to return a bad hypothesis with error at least ϵ, it must be the case that no sample points fell either in the interval $[a, a']$ or the interval $[b', b]$, or both. Otherwise, our algorithm would return a range $[c, d]$ that covers $[a', b']$, and correspondingly the probability our hypothesis would be incorrect on a new input chosen from $\mathcal{D}$ would be at most ϵ.

The probability that a training set of n points does not have any examples from either $[a, a']$ or $[b, b']$ is bounded above by

$$2\left(1 - \frac{\epsilon}{2}\right)^n \leq 2\mathrm{e}^{-\epsilon n/2}.$$

Hence choosing $n \geq 2\ln(2/\delta)/\epsilon$ samples guarantees that the probability of choosing a bad hypothesis is bounded above by δ, and therefore this concept class is PAC learnable.

While the above example of learning intervals demonstrates an infinite concept class that is PAC learnable, the approach to this problem of considering intervals around the maximum and minimum sampled points appears ad hoc. The idea behind this approach, however, can be generalized. Observe that a concept class $\mathcal{C}$ over input set X defines a range space $(X, \mathcal{C})$. We show that the number of examples required to PAC learn a concept class is the same as the number of samples needed to construct an ϵ-net for a range space of VC dimension equal to the VC dimension of the range space defined by the concept class.

Theorem 14.13: *Let $\mathcal{C}$ be a concept class that defines a range space with VC dimension d. For any $0 < \delta, \epsilon \leq 1/2$, there is an*

$$m = O\left(\frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$$

such that $\mathcal{C}$ is PAC learnable with m samples.

Proof: Let X be the ground set of inputs and assume that $c \in \mathcal{C}$ is the correct classification. For any $c' \in \mathcal{C}$, $c' \neq c$ let $\Delta(c', c) = \{x \mid c(x) \neq c'(x)\}$, where $c(x)$ and $c'(x)$ are the labeling functions for c and c' respectively. Let $\Delta(c) = \{\Delta(c', c) \mid c' \in \mathcal{C}\}$. That is, $\Delta(c)$ is a collection of all the possible sets of points of disagreement with the correct classification. The symmetric difference range space with respect to $\mathcal{C}$ and c is $(X, \Delta(c))$. We prove the following lemma about the symmetric difference range space.

Lemma 14.14: *The VC dimension of $(X, \Delta(c))$ is equal to the VC dimension of $(X, \mathcal{C})$.*

Proof: For any set $S \subseteq X$ we define a bijection from the projection of $(X, \mathcal{C})$ on S, denoted by $\mathcal{C}_S$, to the projection of $(X, \Delta(c))$ on S, denoted by $\Delta(c)_S$. The bijection maps each element $c' \cap S \in \mathcal{C}_S$ to $\Delta(c' \cap S, c \cap S) \in \Delta(c)_S$. To show this is a bijection, we first consider two elements $c', c'' \in \mathcal{C}$ with $c' \cap S \neq c'' \cap S$, and show that $\Delta(c' \cap S, c \cap S) \neq \Delta(c'' \cap S, c \cap S)$. If $c' \cap S \neq c'' \cap S$, then there is an element $y \in S$ such that $c'(y) \neq c''(y)$. Without loss of generality, assume that $c'(y) \neq c(y)$ but $c''(y) = c(y)$. In that case $y \in \Delta(c' \cap S, c) \cap S$ but $y \notin \Delta(c'' \cap S, c \cap S)$. Similarly, if for two elements $c', c'' \in \mathcal{C}$ there is an element $y \in S$ such that $\Delta(c' \cap S, c \cap S) \neq \Delta(c'' \cap S, c \cap S)$, then there is an element $y \in S$ such that $c'(y) \neq c''(y)$, so $c' \cap S \neq c'' \cap S$, proving the bijection.

Thus, for any $S \subseteq X$, $|\mathcal{C}_S| = |\Delta(c)_S|$, and S is shattered by $\mathcal{C}$ if and only if it is shattered by $\Delta(c)$. The two range spaces therefore have the same VC dimension. ■

Since the range space $(X, \Delta(c))$ has a VC dimension d, by Theorem 14.8 there is an

$$m = O\left(\frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$$

so that any sample of size m or larger is, with probability at least $1 - \delta$, an ϵ-net for that range space, and therefore has a nonempty intersection with every set $\Delta(c', c)$ that has probability at least ϵ. Thus, with probability at least $1 - \delta$, our training set allows the algorithm to exclude any hypothesis with error probability at least ϵ. ■

We saw in Section 14.2.1 that the VC dimension of the collection of closed intervals on $\mathbb{R}$ is 2. Applying Theorem 14.13 to the problem of learning an interval on the line gives an alternative proof to the result we saw in Section 14.4 that this range space can be learned with $O(\frac{1}{\epsilon} \ln \frac{1}{\delta})$ samples.

14.5. The ϵ-sample Theorem

Recall that an ϵ-sample for a range space $(X, \mathcal{R})$ maintains the relative probability weight of all sets $R \in \mathcal{R}$ within a tolerance of ϵ (Definition 14.6), while an ϵ-net just includes at least one element from each range with total probability at least ϵ. Surprisingly, adding just another $O(1/\epsilon)$ factor to the sample size gives an ϵ-sample, again with probability at least $1 - \delta$. The proof of this result uses the same "double sampling" method as in the proof of the ϵ-net theorem, albeit with a somewhat more complicated argument.

Theorem 14.15: *Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an*

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with probability at least $1 - \delta$.

Proof: Let M be a set of m independent samples from X according to $\mathcal{D}$, and let E_1 be the event that M is not an ϵ-sample for X with respect to the distribution $\mathcal{D}$, i.e.

$$E_1 = \left\{ \exists R \in \mathcal{R} \mid \left| \Pr_{\mathcal{D}}(R) - \frac{|M \cap R|}{|M|} \right| > \epsilon \right\}.$$

We want to show that $\Pr(E_1) \leq \delta$ for a suitable m. We choose a second set T of m independent samples from X according to $\mathcal{D}$, and define E_2 to be the event that some range R is not well approximated by M but is reasonably well approximated by T:

$$E_2 = \left\{ \exists R \in \mathcal{R} \mid \left| \frac{|R \cap M|}{|M|} - \Pr_{\mathcal{D}}(R) \right| > \epsilon \text{ and } \left| \frac{|R \cap T|}{|T|} - \Pr_{\mathcal{D}}(R) \right| \leq \frac{\epsilon}{2} \right\}.$$

Lemma 14.16:

$$\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2).$$

Proof: Clearly the event E_2 is included in the event E_1, thus $\Pr(E_2) \leq \Pr(E_1)$. For the second inequality we again use conditional probability. If E_1 holds, there is some particular R' so that $\left| \frac{|R' \cap M|}{|M|} - \Pr_{\mathcal{D}}(R') \right| > \epsilon$. Therefore,

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr\left(\left| \frac{|R' \cap T|}{|T|} - \Pr_{\mathcal{D}}(R') \right| \leq \frac{\epsilon}{2} \right).$$

Now for a fixed range R' and a random sample T, the random variable $|T \cap R'|$ has a binomial distribution $B(m, \Pr_{\mathcal{D}}(R'))$, and applying the Chernoff bound (Theorem 4.5)

we have

$$\Pr(||T \cap R'| - m \Pr_{\mathcal{D}}(R')| > \epsilon m/2) \leq 2e^{-\epsilon m/12} < 1/2$$

for $m \geq 24/\epsilon$. We conclude

$$\frac{\Pr(E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr\left(\left|\frac{|R' \cap T|}{|T|} - \Pr_{\mathcal{D}}(R')\right| \leq \frac{\epsilon}{2}\right) \geq 1/2. \qquad \blacksquare$$

Next we bound the probability of E_2 by the probability of a larger event E_2':

$$E_2' = \left\{\exists R \in \mathcal{R} \mid ||R \cap T| - |R \cap M|| \geq \frac{\epsilon}{2} m\right\}.$$

To see that $E_2 \subseteq E_2'$, assume that a set R satisfies the conditions of E_2, i.e.

$$||R \cap M| - m \Pr_{\mathcal{D}}(R)| \geq \epsilon m,$$

and

$$||R \cap T| - m \Pr_{\mathcal{D}}(R)| \leq \epsilon m/2.$$

In that case

$$||R \cap M| - m \Pr_{\mathcal{D}}(R)| - ||R \cap T| - m \Pr_{\mathcal{D}}(R)| \geq \epsilon m/2,$$

and by the reverse triangle inequality[2]

$$||R \cap T| - |R \cap M|| \geq ||R \cap M| - m \Pr_{\mathcal{D}}(R)| - ||R \cap T| - m \Pr_{\mathcal{D}}(R)| \geq \epsilon m/2.$$

The event E_2' depends only on the elements in $M \cup T$.

Lemma 14.17:

$$\Pr(E_2) \leq \Pr(E_2') \leq (2m)^d e^{-\epsilon^2 m/8}.$$

Proof: Since M and T are random samples, we can assume that we first choose a random sample of $2m$ elements $Z = z_1, \ldots, z_{2m}$ and then partition it randomly into two sets of size m each. Since Z is a random sample, any partition that is independent of the actual values of the elements generates two random samples. We will use the following partition: for each pair of sampled items z_{2i-1} and z_{2i}, $i = 1, \ldots, m$, with probability $1/2$ (independent of other choices) we place z_{2i-1} in T and z_{2i} in M, otherwise we place z_{2i-1} in M and z_{2i} in T.

For a fixed $R \in \mathcal{R}$, let E_R be the event $\{||R \cap T| - |R \cap M|| \geq \frac{\epsilon}{2} m\}$. To bound the probability of E_R we consider the contribution of the assignment of each pair z_{2i-1}, z_{2i} to the value of $||R \cap T| - |R \cap M||$. If the two items are both in R or the two items are both not in R, the contribution of the pair is 0. If one item is in R and the other is not in R then the contribution of the pair is 1 with probability $1/2$ and -1 with probability

[2] The reverse triangle inequality is simply $|x - y| \geq ||x| - |y||$, which follows easily from the triangle inequality.

1/2. There are no more than m such pairs, so from the Chernoff bound in Theorem 4.7 we can conclude

$$\Pr(E_R) \leq \mathrm{e}^{-\epsilon^2 m/8}.$$

By Theorem 14.1 the projection of $\mathcal{R}$ on Z has no more than $(2m)^d$ ranges. Thus, by the union bound we have

$$\Pr(E_2') \leq (2m)^d \mathrm{e}^{-\epsilon^2 m/8}. \qquad \blacksquare$$

To complete the proof of Theorem 14.15 we show that for

$$m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{2}{\delta}$$

we have

$$\Pr(E_1) \leq 2 \Pr(E_2') \leq 2(2m)^d \mathrm{e}^{-\epsilon^2 m/8} \leq \delta.$$

We remark that this value of m satisfies

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$$

as given in the statement of the theorem; although our explicit bound has a $\ln \frac{64d}{\epsilon^2}$, that term is $O(\ln \frac{d}{\epsilon})$. Equivalently, we require

$$\epsilon^2 m/8 \geq \ln(2/\delta) + d \ln(2m).$$

Clearly it holds that $\epsilon^2 m/16 \geq \ln(2/\delta)$, since $m > \frac{16}{\epsilon^2} \ln \frac{2}{\delta}$. It therefore suffices to show that $\epsilon^2 m/16 \geq d \ln(2m)$ to complete the proof.

Applying Lemma 14.3 with $y = 2m \geq \frac{64d}{\epsilon^2} \ln \frac{64d}{\epsilon^2}$ and $x = \frac{64d}{\epsilon^2}$, we have

$$\frac{4m}{\ln(2m)} \geq \frac{64d}{\epsilon^2},$$

so

$$\frac{\epsilon^2 m}{16} \geq d \ln(2m)$$

as required. $\blacksquare$

Since an ϵ-sample is also an ϵ-net, the lower bound on the sample complexity of ϵ-nets in Theorem 14.11 holds for ϵ-samples. Together with the upper bound of Theorem 14.15, this gives:

Theorem 14.18: *A range space has the uniform convergence property if and only if its VC dimension is finite.*

14.5.1. *Application: Agnostic Learning*

In our discussion of PAC learning in Section 14.4, we assumed that the algorithm is given a concept class $\mathcal{C}$ that includes the correct classification c. That is, there is a

classification that is correct on all items in X, and in particular conforms with all examples in the training set. This assumption does not hold in most applications. First, the training set may have some errors. Second, we may not know any concept class that is guaranteed to include the correct classification and is also simple to represent and compute. In this section we extend our discussion of PAC learning to the case in which the concept class does not necessarily include a perfectly correct classification, which is referred to as the *unrealizable* case or *agnostic* learning. Since the concept class may not have a correct or even close to correct classification, the goal of the the algorithm in this case is to select a classification $c' \in \mathcal{C}$ with an error that is no more than ϵ larger than that of any other classification in $\mathcal{C}$. Formally, let c be the correct classification (which may not be in $\mathcal{C}$). We require the output classification c' to satisfy the following inequality:

$$\Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} \Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.$$

Recall from Section 14.4 that the symmetric difference range space with respect to the concept class $\mathcal{C}$ and the correct classification c is $(X, \Delta(c))$. If the examples in the training set define an $\epsilon/2$-sample for that range space then the algorithm has sufficiently many examples to estimate the error probability of each $c' \in \mathcal{C}$ to within an additive error $\epsilon/2$, and thus can select a classification that satisfies the above requirement.[3] Applying Theorem 14.15, agnostic learning of a concept class with VC dimension d requires $O\left(\min\left(|X|, \frac{d}{\epsilon^2}\ln\frac{d}{\epsilon^2} + \frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)\right)$ samples.

Finally, we state a general characterization of concept classes that are agnostic PAC learnable.

Theorem 14.19: *The following three conditions are equivalent:*

1. *A concept class $\mathcal{C}$ over a domain X is agnostic PAC learnable.*
2. *The range space $(X, \mathcal{C})$ has the uniform convergence property.*
3. *The range space $(X, \mathcal{C})$ has a finite VC dimension.*

14.5.2. *Application: Data Mining*

Data mining involves extracting useful information from raw data. In some cases, such as anomaly detection, one is interested in rare events. Finding such rare events may require a complete analysis of the entire data set that is expensive in both computational time and memory requirements. In other cases, however, the goal of data mining is to detect major patterns or trends in data and ignore random fluctuations. In such settings, analyzing a properly selected sample of the data instead of the entire data set can give an excellent approximation at a fraction of the cost. The crucial question here is how large the sample should be to give a reliable estimate. We give here two examples where using ϵ-samples can provide an answer to this question.

[3] Recall that we are only concerned here with the sampling complexity of the problem. Depending on the particular concept class, the computation cost may not be practically feasible.

Example: Estimating dense neighborhoods

Assume that we are given a large set of n points in the plane and we need to answer a sequence of queries of the form "what fraction of the points are at distance at most r from point (x, y)?", for arbitrary values of (x, y) and r. Estimates of this kind are used by businesses to determine where to locate new stores or other resources. For example, if points represent home locations for customers, query locations (x, y) might represent possible locations for a bank to place an automated teller machine, in which case quick estimates of how many customers are near the location would be useful for planning purposes.

We can answer each query by scanning the entire set of n points. Alternatively, we can define a range space $(\mathbb{R}^2, \mathcal{R})$, where $\mathcal{R}$ includes, for each pair $(x, y) \in \mathbb{R}^2$ and $r \in \mathbb{R}^+$, the set of all the points inside the disk of radius r centered at (x, y). Since the VC dimension of the set of all disks on the plane is 3 (see Exercise 14.6), we can sample a random set of $m = O\left(\frac{1}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$ points and give fast approximate answers to all the queries by scanning only the sample.

Generating a random sample may require an initial scan of all the n points, but we need to execute it only once. The ϵ-sample theorem guarantees that with probability at least $1 - \delta$, the answers to *all* of the queries are within ϵ of the correct value. Furthermore, since the ϵ-sample estimates all possible disks, we could also use it for other purposes, such as approximately identifying the k densest disks.

Example: Mining frequent itemsets

Consider a supermarket that wants to design discounts for customers based on buying a collection of items. In this case, the supermarket is interested not only in what are the most frequent items purchased, but also in what sets of items are most frequently bought together. This problem arises in many settings, and is commonly referred to as the problem of mining frequent itemsets. Formally, we can describe the problem as follows: we are given a set of items $\mathcal{I}$ and a collection of transactions $\mathcal{T}$, where a transaction is a subset of $\mathcal{I}$. We are interested in sets of items that appear in many transactions, where what is meant by many transactions can depend on the setting. We might use a threshold, or a percentage of transactions.

Mining frequent itemsets is challenging to accomplish efficiently, both because the number of customer transactions is usually large, and because it takes significant memory to store all possible frequent itemsets. Even if one limits the problem to itemsets of size up to k, there are $\binom{|\mathcal{I}|}{k}$ possible itemsets, which grows large even for small k. All known exact solutions to this problem require either several passes over the data or significant storage or both to store candidate frequent itemsets and their counts. On the other hand, solving the problem on a relatively small sample can give effective results much more efficiently.

A natural goal would be to make sure we find all sufficiently frequent itemsets and discard all sufficiently infrequent itemsets. There might be some itemsets that are ambiguously in between the thresholds we set for frequent and infrequent itemsets, and that could therefore be characterized either way. Suppose that we want to correctly characterize all sets with frequency greater than θ as frequent and all sets with

frequency less than $\theta - \epsilon$ as infrequent; sets with frequency $[\theta - \epsilon, \theta]$ would be in the ambiguous range. How many transactions do we need to sample?

Our goal is to approximate the true frequency of each set within an additive error of $\epsilon/2$. Then we can treat all sets with frequency at least $\theta - \epsilon/2$ as frequent itemsets and all sets with frequency less than $\theta - \epsilon/2$ as infrequent itemsets, ensuring that we correctly characterize sets with frequency greater than θ and sets with frequency less than $\theta - \epsilon$.

If all transactions have size at most ℓ, then there are $O(|\mathcal{I}|^\ell)$ different itemsets that could be frequent. Applying a Chernoff bound and a union bound would require a sample of size $\Omega\left(\frac{\theta}{\epsilon^2}\left(\ell \ln |\mathcal{I}| + \ln \frac{1}{\delta}\right)\right)$. In practice, $\ell << |\mathcal{I}|$. In such a case an ϵ-sample can give a significantly better bound. (Although, strictly speaking, here we need an $(\epsilon/2)$-sample.)

For each subset $s \subseteq \mathcal{I}$, let $T(s) = \{t \in \mathcal{T} \text{ and } s \subseteq t\}$ denote the collection of all transactions in the data set that include s. Let $\mathcal{R} = \{T(s) \mid s \subseteq \mathcal{I}\}$, and consider the range space $(\mathcal{T}, \mathcal{R})$. We would like to bound the VC dimension of this range space by a parameter that can be evaluated in one pass over the data (say when the data is first loaded to the system). We first observe that the VC dimension is bounded by ℓ, the maximum size of any transaction in the data set. Indeed, a transaction of size q has 2^q subsets and is therefore included in no more than 2^q ranges. Since no transaction can belong to more than 2^ℓ ranges, no set of more than ℓ transactions can be shattered. Thus, by Theorem 14.15, with probability at least $1 - \delta$, a sample of size

$$O\left(\frac{\ell}{\epsilon^2} \ln \frac{\ell}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right) \tag{14.1}$$

can guarantee that all itemsets are accurately determined to within $\epsilon/2$ of their true proportion with probability at least $1 - \delta$, and thus is sufficient for identifying all the frequent itemsets. A better bound is proven in Exercise 14.12.

14.6. Rademacher Complexity

Rademacher complexity is an alternative approach for computing sample complexity. Unlike the VC-dimension based bounds, which were distribution independent, the Rademacher complexity bounds depend on the training set distribution, and thus can give better bounds for specific input distributions. Furthermore, the Rademacher complexity can, in principle, be estimated from the training set, allowing for strong bounds derived from a sample itself. Another advantage of Rademacher complexity is that it can be applied to the estimation of any function, not just 0–1 classification functions. (There are, to be clear, generalizations of VC dimensions to non-binary function.)

To motivate the definition of Rademacher averages, let us start with the binary classification setting we used in section 14.1 and then generalize. We have a training set $(x_1, c(x_1)), \ldots, (x_m, c(x_m))$ where $x_i \in U$ and $c(x_i) \in \{-1, 1\}$, and a set of possible hypotheses $h \in \mathcal{C}$ where each h is a function from the universe U to $\{-1, 1\}$. The *training error* of a hypothesis on the training set is the fraction of samples where the

hypothesis disagrees with the given classification. Formally,

$$\hat{err}(h) = \frac{1}{m}|\{i : h(x_i) \neq c(x_i), 1 \leq i \leq m\}|.$$

Now we make use of the fact that, because $h(x_i)$ and $c(x_i)$ take on values in $\{-1, 1\}$,

$$\frac{1 - c(x_i)h(x_i)}{2} = \begin{cases} 0 & \text{if } c(x_i) = h(x_i), \\ 1 & \text{if } c(x_i) \neq h(x_i). \end{cases}$$

Hence we can write

$$\begin{aligned} \hat{err}(h) &= \frac{1}{m}\sum_{i=1}^{m} \frac{1 - c(x_i)h(x_i)}{2} \\ &= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} c(x_i)h(x_i). \end{aligned}$$

The expression $\frac{1}{m}\sum_{i=1}^{m} c(x_i)h(x_i)$ represents the correlation between c and h; if c and h always agree, the value of the expression is 1, and if they alway disagree, the value is -1. The hypothesis that minimizes the training error is the hypothesis that maximizes the correlation.

Now, given a collection of sample points x_i, $1 \leq i \leq m$, we consider how well our class of possible hypotheses $\mathcal{C}$ can align with all possible classifications of these sample points. To consider all possible classifications, we use the *Rademacher variables*: m independent random variables, $\sigma = (\sigma_1, \ldots, \sigma_m)$, with $\Pr(\sigma_i = -1) = \Pr(\sigma_i = 1) = 1/2$. The hypothesis that aligns best with fixed values of the Rademacher variables σ is then the one that maximizes the value

$$\frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i),$$

and our training error is

$$\frac{1}{2} - \max_{h \in \mathcal{C}} \frac{1}{2m}\sum_{i=1}^{m} \sigma_i h(x_i).$$

To consider all possible sample points, we consider the expectation over all possible outcomes for σ, or

$$\mathbf{E}_\sigma \max_{h \in \mathcal{C}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i). \tag{14.2}$$

This expression corresponds intuitively to how expressive our class of hypotheses $\mathcal{C}$ is. For example, if $\mathcal{C}$ consisted of just a single hypothesis h, the expectation would be 0, as $h(x_i) = \sigma_i$ with probability $1/2$ for any randomly chosen σ. On the other hand, if $\mathcal{C}$ shatters the set $\{x_1, x_2, \ldots, x_m\}$, then the expectation would be 1, as there would be some $h \in \mathcal{C}$ so that $h(x_i) = \sigma_i$ for all i for each possible randomly chosen σ. In this particular setting, the expectation is always between 0 and 1, and intuitively a higher number corresponds to a more expressive set of hypotheses.

To move to a more general definition of Rademacher averages, instead of thinking of sets of hypotheses, we consider a set of real-valued functions $\mathcal{F}$, where the inputs to the

function are defined according to a probability space with distribution $\mathcal{D}$. Hence, for $f \in \mathcal{F}$, when we refer to $\mathbf{E}[f]$, this would correspond to $\mathbf{E}[f(Z)]$ where Z is a random variable with distribution $\mathcal{D}$. We generalize the expectation (14.2) as follows.

Definition 14.10: *The* empirical Rademacher average *of a set of functions $\mathcal{F}$ with respect to a sample $S = \{z_1, \ldots, z_m\}$, is defined as*

$$\tilde{R}_m(\mathcal{F}, S) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where the expectation is taken over the distribution of the Rademacher variables $\sigma = (\sigma_1, \ldots, \sigma_m)$.

We remark that we use sup instead of max since we are dealing with a family of real-valued functions, so the maximum technically may not exist.

For a fixed assignment of values to the Rademacher variables the value of $\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)$ represents the best correlation between any function in $\mathcal{F}$ and the vector $(\sigma_1, \ldots, \sigma_m)$, generalizing the correlation for binary classifications. The empirical Rademacher average therefore measures how well one can correlate random partitions of the sample with some function in the set $\mathcal{F}$, which provides a measure of how expressive the set is. We therefore use the terms empirical Rademacher average and empirical Rademacher complexity interchangeably (both terms are used in the literature).

Now let us look at the empirical Rademacher average in a different way. For large m, an average $\frac{1}{m} \sum_{i=1}^{m} f(z_i)$ over a random sample $S = \{z_1, \ldots, z_m\}$, should provide a good approximation to $\mathbf{E}[f]$. Multiplying by the Rademacher variables, the expression $\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)$ corresponds to splitting the sample S into two subsamples, corresponding to the values of i where $\sigma_i = 1$ and the values of i where $\sigma_i = -1$. If S is a random sample then the expression is similar to the difference between the average of the two random subsamples, and hence the expectation

$$\mathbf{E}_\sigma \left[\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

should be small. Finally, the empirical Rademacher complexity

$$\tilde{R}_m(\mathcal{F}, S) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

considers the supremum of this expectation over all functions in $\mathcal{F}$. Intuitively, if the empirical Rademacher average with respect to a sample of size m is small, then we expect m to be sufficiently large for a sample to provide a good estimate for all functions in $\mathcal{F}$. We formulate and prove this intuition in Theorem 14.20.

To remove the dependency on a particular sample we can take an expectation over the distribution of all samples S of size m, where the samples are taken from the distribution $\mathcal{D}$.

Definition 14.11: *The* Rademacher average *of $\mathcal{F}$ is defined as*

$$R_m(\mathcal{F}) = \mathbf{E}_S[\tilde{R}_m(\mathcal{F}, S)] = \mathbf{E}_S \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where the expectation over S corresponds to samples of size m from a given distribution $\mathcal{D}$.

We similarly use the terms Rademacher average and Rademacher complexity interchangeably.

14.6.1. *Rademacher Complexity and Sample Error*

A key property of the Rademacher complexity of a set of functions $\mathcal{F}$ is that it bounds the expected maximum error in estimating the mean of any function $f \in \mathcal{F}$ using a sample.

Let $\mathbf{E}_\mathcal{D}[f(z)]$ be the true mean of f with respect to distribution $\mathcal{D}$. The estimate of $\mathbf{E}_\mathcal{D}[f(z)]$ using the sample $S = \{z_1, \ldots, z_m\}$ is $\frac{1}{m}\sum_{i=1}^{m} f(z_i)$. The expected maximum error, averaged over all samples of size m from $\mathcal{D}$, is given by

$$\mathbf{E}_S \left[\sup_{f \in \mathcal{F}} \left(\mathbf{E}_\mathcal{D}[f(z)] - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right) \right].$$

The following theorem bounds this error in terms of the Rademacher complexity of $\mathcal{F}$.

Theorem 14.20:

$$\mathbf{E}_S \left[\sup_{f \in \mathcal{F}} \left(\mathbf{E}_\mathcal{D}[f(z)] - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right) \right] \leq 2R_m(\mathcal{F}).$$

Proof: Pick a second sample $S' = \{z'_1, \ldots, z'_m\}$.

$$\begin{aligned}
&\mathbf{E}_S \left[\sup_{f \in \mathcal{F}} \left(\mathbf{E}_\mathcal{D}[f(z)] - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right) \right] \\
&= \mathbf{E}_S \left[\sup_{f \in \mathcal{F}} \left(\mathbf{E}_{S'} \frac{1}{m} \sum_{i=1}^{m} f(z'_i) - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right) \right] \\
&\leq \mathbf{E}_{S,S'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(z'_i) - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right) \right] \\
&= \mathbf{E}_{S,S',\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(z_i) - f(z'_i)) \right) \right] \\
&\leq \mathbf{E}_{S,\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right] + \mathbf{E}_{S',\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z'_i) \right] \\
&= 2R_m(\mathcal{F}).
\end{aligned}$$

The first equality holds because the expectation from the sample S' is the expectation of f. The first inequality, in which the order of the expectation with respect to S' with the operation $\sup_{f\in\mathcal{F}}$ is interchanged, follows from Jensen's inequaliy (Theorem 2.4), and the fact that supremum is a convex function. For the second equality, we use the fact that multiplying $f(z_i) - f(z_i')$ by a Rademacher variable σ_i does not change the expectation of the sum. If $\sigma_i = 1$ there is clearly no change, and if $\sigma_i = -1$ this is equivalent to switching z_i and z_i' between the two samples, which does not change the expectation. For the second inequality, we use that σ_i and $-\sigma_i$ have the same distribution, so we can change the sign to simplify the expression. ■

Next we show that for bounded functions the Rademacher complexity is well approximated by the empirical Rademacher complexity, and the estimation error is well approximated by twice the Rademacher complexity, thereby obtaining a probabilistic bound on the estimation error of any bounded function in $\mathcal{F}$ from a sample.

Theorem 14.21: *Let $\mathcal{F}$ be a set of functions such that for any $f \in \mathcal{F}$ and for any two values x and y in the domain of f, $|f(x) - f(y)| \leq c$ for some constant c. Let $R_m(\mathcal{F})$ be the Rademacher complexity, and $\tilde{R}_m(\mathcal{F}, S)$ the empirical Rademacher complexity of the set $\mathcal{F}$, with respect to a random sample $S = \{z_1, \ldots, z_m\}$ of size m from a distribution $\mathcal{D}$.*

(1) *For any $\epsilon \in (0, 1)$,*

$$\Pr(|\tilde{R}_m(\mathcal{F}, S) - R_m(\mathcal{F})|) \geq \epsilon) \leq 2\mathrm{e}^{-2m\epsilon^2/c^2}.$$

(2) *For all $f \in \mathcal{F}$ and $\epsilon \in (0, 1)$,*

$$\Pr\left(\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \geq 2\tilde{R}_m(\mathcal{F}, S) + 3\epsilon\right) \leq 2\mathrm{e}^{-2m\epsilon^2/c^2}.$$

Proof: To prove the first part of the theorem we observe that $\tilde{R}_m(\mathcal{F}, S)$ is a function of m random variables, $z_1, \ldots, z_m$, and any change in one of these variables can change the value of $\tilde{R}_m(\mathcal{F}, S)$ by no more than c/m. Since $\mathbf{E}_S[\tilde{R}_m(\mathcal{F}, S)] = R_m(\mathcal{F})$ we can apply Theorem 13.7 to obtain

$$\Pr(|\tilde{R}_m(\mathcal{F}, S) - R_m(\mathcal{F}|) \geq \epsilon) \leq 2\mathrm{e}^{-2m\epsilon^2/c^2}.$$

To prove the second part, we observe that $\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)$ is a function of $z_1, \ldots, z_m$, and a change in one of the z_i changes the value of that function by no more than c/m. Applying a one-sided form of Theorem 13.7 we have

$$\Pr\left(\left(\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right) - \mathbf{E}_S\left[\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right] \geq \epsilon\right) \leq \mathrm{e}^{-2m\epsilon^2/c^2}.$$

We now apply the bound in Theorem 14.20,

$$\mathbf{E}_S\left[\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right] \leq 2R_m(\mathcal{F}),$$

to obtain,

$$\Pr\left(\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \geq 2R_m(\mathcal{F}) + \epsilon\right) \leq \mathrm{e}^{-2m\epsilon^2/c^2}. \tag{14.3}$$

From the first part of the theorem we know that $R_m(\mathcal{F}) \leq \tilde{R}_m(\mathcal{F}, S) + \epsilon$ with probability at least $1 - \mathrm{e}^{-2m\epsilon^2/c^2}$. Combining this with Eqn. (14.3), we have the second part of the theorem,

$$\Pr\left(\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \geq 2\tilde{R}_m(\mathcal{F}, S) + 3\epsilon\right) \leq 2\mathrm{e}^{-2m\epsilon^2/c^2}. \qquad \blacksquare$$

14.6.2. *Estimating the Rademacher Complexity*

While the Rademacher complexity can, in principle, be computed from a sample, in practice it is often hard to compute the expected supremum over a large (or even infinite) set of functions. Massart's theorem provides a bound that is often easy to compute for finite sets of functions.

Theorem 14.22 [Massart's theorem]: *Assume that $|\mathcal{F}|$ is finite. Let $S = \{z_1, \ldots, z_m\}$ be a sample, and let*

$$B = \max_{f\in\mathcal{F}} \left(\sum_{i=1}^{m} f^2(z_i)\right)^{\frac{1}{2}}$$

then

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2\ln|\mathcal{F}|}}{m}.$$

Proof: For any $s > 0$,

$$\mathrm{e}^{sm\tilde{R}_m(\mathcal{F},S)} = \mathrm{e}^{s\mathbf{E}_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)]},$$

where the expectation is taken over the assignments of the Rademacher variables $\sigma = (\sigma_1, \ldots, \sigma_m)$.

By Jensen's inequality (Theorem 2.4),

$$\begin{aligned}
\mathrm{e}^{s\mathbf{E}_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)]} &\leq \mathbf{E}_\sigma\left[\mathrm{e}^{s\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)}\right] \\
&= \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\mathrm{e}^{\sum_{i=1}^{m} s\sigma_i f(z_i)}\right)\right] \\
&\leq \sum_{f\in\mathcal{F}} \mathbf{E}_\sigma\left[\left(\mathrm{e}^{\sum_{i=1}^{m} s\sigma_i f(z_i)}\right)\right] \\
&= \sum_{f\in\mathcal{F}} \mathbf{E}_\sigma\left[\prod_{i=1}^{m} \mathrm{e}^{s\sigma_i f(z_i)}\right] \\
&= \sum_{f\in\mathcal{F}} \prod_{i=1}^{m} \mathbf{E}_\sigma\left[\mathrm{e}^{s\sigma_i f(z_i)}\right].
\end{aligned}$$

Here the first line follows from Jensen's inequality, and the second line is just a rearrangement of terms. The third line bounds the supremum by a summation, which is possible since all the terms are positive. The fourth line changes the sum in the exponent to a product, and the last line arises from the independence of the sample values.

Since $\mathbf{E}[\sigma_i f(z_i)] = 0$ and $-f(z_i) \leq \sigma_i f(z_i) \leq f(z_i)$, we can apply Hoeffding's Lemma (Lemma 4.13) to obtain

$$\mathbf{E}\left[e^{s\sigma_i f(z_i)}\right] \leq e^{s^2(2f(z_i))^2/8} = e^{s^2 f(z_i)^2/2}.$$

Thus,

$$\begin{aligned} e^{sm\tilde{R}_m(\mathcal{F},S)} &= e^{s\mathbf{E}[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)]} \\ &\leq \sum_{f\in\mathcal{F}} \prod_{i=1}^m e^{s^2 f(z_i)^2/2} \\ &= \sum_{f\in\mathcal{F}} e^{s^2/2 \sum_{i=1}^m f(z_i)^2} \\ &\leq |\mathcal{F}| e^{(s^2B^2)/2}. \end{aligned}$$

Hence, for any $s > 0$,

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{1}{m}\left(\frac{\ln|\mathcal{F}|}{s} + \frac{sB^2}{2}\right).$$

Setting $s = \frac{\sqrt{2\ln|\mathcal{F}|}}{B}$ yields

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2\ln|\mathcal{F}|}}{m}.$$ ■

14.6.3. *Application: Agnostic Learning of a Binary Classification*

Let $\mathcal{C}$ be a binary concept class defined on a domain X, and let $\mathcal{D}$ be a probability distribution on X. For each $x \in X$ let $c(x)$ be the correct classification of x. For each hypothesis $h \in C$ we define a function $f_h(x)$ by

$$f_h(x) = \begin{cases} 1 & \text{if } h(x) = c(x) \\ -1 & \text{otherwise.} \end{cases}$$

Let $\mathcal{F} = \{f_h \mid h \in \mathcal{C}\}$. Our goal is to find $h' \in \mathcal{C}$ such that with probability at least $1 - \delta$

$$\mathbf{E}[f_{h'}] \geq \sup_{f_h\in\mathcal{F}} \mathbf{E}[f_h] - \epsilon.$$

Let S be a sample of size m. We apply Theorem 14.22 to bound the empirical Rademacher average $\mathcal{F}$ with respect to S. Since the functions in $\mathcal{F}$ take on only the values -1 and 1,

$$B = \max_{f\in\mathcal{F}} \left(\sum_{i=1}^m f^2(z_i)\right)^{\frac{1}{2}} = \sqrt{m},$$

and for a finite $\mathcal{F}$

$$\tilde{R}_m(\mathcal{F}, S) \leq \sqrt{\frac{2 \ln |\mathcal{F}|}{m}}.$$

Next we express this bound in terms of the VC dimension of the concept class $\mathcal{C}$. Each function $f_h \in \mathcal{F}$ corresponds to a hypothesis $h \in \mathcal{C}$. Let d be the VC dimension of $\mathcal{C}$. The projection of the range space $(X, \mathcal{C})$ on a sample of size m has no more than m^d different sets, as we know from Theorem 14.1. Thus, the set of different functions we need to consider is bounded by m^d, and hence

$$\tilde{R}_m(\mathcal{F}, S) \leq \sqrt{\frac{2d \ln m}{m}}.$$

The bound on $\tilde{R}_m(\mathcal{F}, S)$ in conjunction with Theorem 14.21 can be used to obtain an alternative bound on the sample complexity of agnostic learning, similar to the bound found in Section 14.5.1. The details are considered in Exercise 14.15. However, for specific distributions, the projection of $(X, \mathcal{C})$ on the training set can be significantly smaller, yielding a smaller Rademacher complexity and smaller sample complexity.

14.7. Exercises

Exercise 14.1: Consider a range space $(X, \mathcal{C})$ where $X = \{1, 2, \ldots, n\}$ and $\mathcal{C}$ is the set of all subsets of X of size k for some $k < n$. What is the VC dimension of $\mathcal{C}$?

Exercise 14.2: Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all axis-aligned rectangles in $\mathbb{R}^2$. That is, $c \in \mathcal{C}$ if for some $x_0 < x_1$ and $y_0 < y_1$, $c = \{(x, y) \in \mathbb{R}^2 \mid x_0 \leq x \leq x_1 \text{ and } y_0 \leq y \leq y_1\}$.

(a) Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 4. You should show both a set of four points that can be shattered, and show that no larger set can be shattered.

(b) Construct and analyze a PAC learning algorithm for the concept class of all axis-aligned rectangles in $\mathbb{R}^2$.

Exercise 14.3: Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all axis-aligned squares in $\mathbb{R}^2$. Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 3.

Exercise 14.4: Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all squares (that need not be axis-aligned) in $\mathbb{R}^2$. Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 5.

Exercise 14.5: Consider a range space $(\mathbb{R}^3, \mathcal{C})$ of all axis-aligned rectangular boxes in $\mathbb{R}^3$. Find the VC dimension of $(\mathbb{R}^3, \mathcal{C})$; you should show both the largest number of points that can be shattered, and show that no larger set can be shattered.

Exercise 14.6: Prove that the VC dimension of the collection of all closed disks on the plane is 3.

Exercise 14.7: Prove that the VC dimension of the range space $(\mathbb{R}^d, \mathcal{R})$, where $\mathcal{R}$ is the set of all half-spaces in $\mathbb{R}^d$, is at least $d + 1$, by showing that the set consisting of the origin $(0, 0, \ldots, 0)$ and the d unit points $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 1)$ is shattered by $\mathcal{R}$.

Exercise 14.8: Let $S = (X, R)$ and $S' = (X, R')$ be two range spaces. Prove that if $R' \subseteq R$ then the VC dimension of S' is no larger than the VC dimension of S.

Exercise 14.9: Show that for $n \geq 2d$ and $d \geq 1$ the growth function satisfies

$$\mathcal{G}(d, n) = \sum_{i=0}^{d} \binom{n}{i} \leq 2 \left(\frac{ne}{d}\right)^d.$$

Exercise 14.10: Use the bound of Exercise 14.9 to improve the result of Theorem 14.4 to show the VC dimension of the range space $(X, \mathcal{R}^f)$ is $O(kd \ln k)$.

Exercise 14.11: Use the bound of Exercise 14.9 to improve the result of Theorem 14.8 to show that there is an

$$m = O\left(\frac{d}{\epsilon} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m suffices to obtain the required ϵ-net with probability at least $1 - \delta$. (*Hint*: Use Lemma 14.3 with $x = O(\frac{1}{\epsilon})$ and $y = \frac{2m}{d}$.)

Exercise 14.12: **(a)** Improve the result in Eqn. (14.1) by showing that the VC dimension of the frequent-itemsets range space is bounded by the maximum number q such that the data set has q different transactions all of size at least q.

(b) Show how to compute an upper bound on the number q defined in (a) in one pass over the data.

Exercise 14.13: Prove Theorem 14.11 using the following hints. Let (X, R) be a range space with VC dimension d. Let $Y = \{y_1, \ldots, y_d\} \subseteq X$ be a set of d elements that is shattered by R. Define a probability distribution $\mathcal{D}$ on R as follows: $\Pr(y_1) = 1 - 16\epsilon$, $\Pr(y_2) = \Pr(y_3) = \cdots = Pr(y_d) = 16\epsilon/(d - 1)$, and all other elements have probability 0. Consider a sample of size $m = (d - 1)/(64\epsilon)$. Show that with probability at least $1/2$ the sample does not include at least half of the elements in $\{y_2, \ldots, y_d\}$. Conclude that with probability $\delta \geq 1/2$ the output classification has error at least ϵ.

Exercise 14.14: Given a set of functions $\mathcal{F}$ and constants $a, b \in \mathbb{R}$, consider the set of functions

$$\mathcal{F}_{a,b} = \{af + b \mid f \in \mathcal{F}\}.$$

Let $R_m()$ and $\tilde{R}_m()$ denote the Rademacher complexity and the empirical Rademacher complexity, respectively. Prove that

(a) $\tilde{R}_m(\mathcal{F}_{a,b}, S) = |a|\tilde{R}_m(\mathcal{F}, S)$,
(b) $R_m(\mathcal{F}_{a,b}) = |a|R_m(\mathcal{F})$.

Exercise 14.15: We apply Theorem 14.21 to compute a bound on the sample complexity of agnostic learning a binary classification. Assume a concept class with VC dimension d and a sample size m.

(a) Find a sample size m_1 such that the Empirical Rademacher Average of the corresponding set of functions is at most $\epsilon/4$.
(b) Use Theorem 14.21 to find a sample size m such that with probability at least $1 - \delta$ the expectation of all the functions are estimated within error ϵ.
(c) Compare your bound to the result obtained in Section 14.5.1.