VAPNIK–CHERVONENKIS CLASSES

MAXIM RAGINSKY

A key result on the ERM algorithm, proved in the previous lecture, was that

\[ P(\hat{f}_n) \leq L^*(F) + 4\mathbb{ER}_n(F(Z^n)) + \sqrt{\frac{2\log(1/\delta)}{n}} \]

with probability at least 1 – \( \delta \). The quantity \( R_n(F(Z^n)) \) appearing on the right-hand side of the above bound is the Rademacher average of the random set

\[ F(Z^n) = \{(f(Z_1), \ldots, f(Z_n)) : f \in F\}, \]

often referred to as the projection of \( F \) onto the sample \( Z^n \). From this we see that a sufficient condition for the ERM algorithm to produce near-optimal hypotheses with high probability is that the expected Rademacher average

\[ \mathbb{ER}_n(F(Z^n)) = \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \]

where the \( \tilde{O}(\cdot) \) notation indicates that the bound holds up to polylogarithmic factors in \( n \), i.e., there exists some positive polynomial function \( p(\cdot) \) such that

\[ \mathbb{ER}_n(F(Z^n)) \leq O\left(\sqrt{\frac{\log n}{n}}\right). \]

Hence, a lot of effort in statistical learning theory is devoted to obtaining tight bounds on \( \mathbb{ER}_n(F(Z^n)) \).

One way to guarantee an \( \tilde{O}(1/\sqrt{n}) \) bound on \( \mathbb{ER}_n \) is if the “effective size” of the random set \( F(Z^n) \) is finite and grows polynomially with \( n \). Then the Finite Class Lemma will tell us that

\[ R_n(F(Z^n)) = O\left(\sqrt{\frac{\log n}{n}}\right). \]

In general, a reasonable notion of “effective size” is captured by various covering numbers (see, e.g., the lecture notes by Mendelson [Men03] or the recent monograph by Talagrand [Tal05] for detailed expositions of the relevant theory). In this lecture, we will look at a simple combinatorial notion of effective size for classes of binary-valued functions. This particular notion has originated with the work of Vapnik and Chervonenkis [VC71], and was historically the first such notion to be introduced into statistical learning theory. It is now known as the Vapnik–Chervonenkis (or VC) dimension.

1. Vapnik–Chervonenkis dimension: definition

**Definition 1.** Let \( C \) be a class of (measurable) subsets of some space \( Z \). We say that a finite set \( S = \{z_1, \ldots, z_n\} \subset Z \) is shattered by \( C \) if for every subset \( S' \subseteq S \) there exists some \( C \in C \) such that \( S' = S \cap C \).

In other words, \( S = \{z_1, \ldots, z_n\} \) is shattered by \( C \) if for any binary \( n \)-tuple \( b = (b_1, \ldots, b_n) \in \{0,1\}^n \) there exists some \( C \in C \) such that

\[ (1_{\{z_1 \in C\}}, \ldots, 1_{\{z_n \in C\}}) = b \]

*Date: February 23, 2011.*
or, equivalently, if
\[ \{ (1_{\{z_1 \in C\}} \ldots, 1_{\{z_n \in C\}}) : C \in \mathcal{C} \} = \{0, 1\}^n, \]
where we consider any two \( C_1, C_2 \in \mathcal{C} \) as equivalent if \( 1_{\{z_1 \in C_1\}} = 1_{\{z_i \in C_2\}} \) for all \( 1 \leq i \leq n \).

**Definition 2.** The Vapnik-Chervonenkis dimension (or the VC dimension) of \( \mathcal{C} \) is
\[ V(\mathcal{C}) \triangleq \max \left\{ n \in \mathbb{N} : \exists S \subseteq \mathbb{Z} \text{ such that } |S| = n \text{ and } S \text{ is shattered by } \mathcal{C} \right\}. \]

If \( V(\mathcal{C}) < \infty \), we say that \( \mathcal{C} \) is a VC class (of sets).

We can express the VC dimension in terms of shatter coefficients of \( \mathcal{C} \): Let
\[ S_n(\mathcal{C}) \triangleq \sup_{S \subseteq \mathbb{Z}, |S| = n} |\{S \cap C : C \in \mathcal{C}\}| \]
denote the \( n \)th shatter coefficient of \( \mathcal{C} \), where for each fixed \( S \) we consider any two \( C_1, C_2 \in \mathcal{C} \) as equivalent if \( S \cap C_1 = S \cap C_2 \). Then
\[ V(\mathcal{C}) = \max \left\{ n \in \mathbb{N} : S_n(\mathcal{C}) = 2^n \right\}. \]

The VC dimension \( V(\mathcal{C}) \) may be infinite, but it is always well-defined. This follows from the following lemma:

**Lemma 1.** If \( S_n(\mathcal{C}) < 2^n \), then \( S_m(\mathcal{C}) < 2^m \) for all \( m > n \).

**Proof.** Suppose \( S_n(\mathcal{C}) < 2^n \). Consider any \( m > n \). We will suppose that \( S_m(\mathcal{C}) = 2^m \) and derive a contradiction. Let us choose an arbitrary binary \( n \)-tuple \( b = (b_1, \ldots, b_n) \). Moreover, choose an arbitrary \( S = \{z_1, \ldots, z_n\} \subseteq \mathbb{Z} \) and some other \( T = \{z'_1, \ldots, z'_{m-n}\} \subseteq \mathbb{Z} \) with \( S \cap T = \emptyset \). By our assumption that \( S_m(\mathcal{F}) = 2^m \), given
\[ b' = (b_1, \ldots, b_n, 0, \ldots, 0) \]
and \( S' = S \cup T = \{z_1, \ldots, z_n, z'_1, \ldots, z'_{m-n}\} \), we can find some \( C \in \mathcal{C} \) such that
\[ (1_{\{z_1 \in C\}}, \ldots, 1_{\{z_n \in C\}}, 1_{\{z'_1 \in C\}}, \ldots, 1_{\{z'_{m-n} \in C\}}) = (b_1, \ldots, b_n, 0, \ldots, 0). \]
From (1) it immediately follows that
\[ (1_{\{z_1 \in C\}}, \ldots, 1_{\{z_n \in C\}}) = (b_1, \ldots, b_n) = b. \]
Since \( b = (b_1, \ldots, b_n) \) and \( S = \{z_1, \ldots, z_n\} \) were arbitrary, we see from (2) that \( S_n(\mathcal{C}) = 2^n \). This contradicts our assumption that \( S_n(\mathcal{C}) < 2^n \), so we conclude that \( S_m(\mathcal{C}) < 2^m \) whenever \( m > n \) and \( S_n(\mathcal{F}) < 2^n \). \( \square \)

There is a one-to-one correspondence between binary-valued functions \( f : \mathbb{Z} \to \{0, 1\} \) and subsets of \( \mathbb{Z} \):
\[ \forall f : \mathbb{Z} \to \{0, 1\} \text{ let } \mathcal{C}_f \triangleq \{z : f(z) = 1\} \text{ and } \forall C \subseteq \mathbb{Z} \text{ let } f_C \triangleq 1_{\{C\}}. \]
Thus, we can extend the concept of shattering, as well as the definition of the VC dimension, to any class \( \mathcal{F} \) of functions \( f : \mathbb{Z} \to \{0, 1\} \):
Definition 3. Let \( \mathcal{F} \) be a class of functions \( f : Z \to \{0,1\} \). We say that a finite set \( S = \{z_1, \ldots, z_n\} \subset Z \) is shattered by \( \mathcal{F} \) if it is shattered by the class

\[
\mathcal{C}_\mathcal{F} \triangleq \{ \mathbf{1}_{\{f=1\}} : f \in \mathcal{F} \},
\]

where \( \mathbf{1}_{\{f=1\}} \) is the indicator function of the set \( C_f \triangleq \{ z \in Z : f(z) = 1 \} \). The \( n \)th shatter coefficient of \( \mathcal{F} \) is \( S_n(\mathcal{F}) = S_n(\mathcal{C}_\mathcal{F}) \), and the VC dimension of \( \mathcal{F} \) is defined as \( V(\mathcal{F}) = V(\mathcal{C}_\mathcal{F}) \).

In light of these definitions, we can equivalently speak of the VC dimension of a class of sets or a class of binary-valued functions.

2. Examples of Vapnik–Chervonenkis classes

2.1. Semi-infinite intervals. Let \( Z = \mathbb{R} \) and take \( \mathcal{C} \) to be the class of all intervals of the form \((−\infty,t]\) as \( t \) varies over \( \mathbb{R} \). We will prove that \( V(\mathcal{C}) = 1 \). In view of Lemma 1, it suffices to show that (1) any one-point set \( S = \{a\} \) is shattered by \( \mathcal{C} \), and (2) no two-point set \( S = \{a,b\} \) is shattered by \( \mathcal{C} \).

Given \( S = \{a\} \), choose any \( t_1 < a \) and \( t_2 > a \). Then \((−\infty,t_1]\cap S = \emptyset\) and \((−\infty,t_2]\cap S = S\). Thus, \( S \) is shattered by \( \mathcal{C} \). This holds for every one-point set \( S \), and therefore we have proved (1). To prove (2), let \( S = \{a,b\} \) and suppose, without loss of generality, that \( a < b \). Then there exists no \( t \in \mathbb{R} \) such that \((−\infty,t]\cap S = \{b\}\). This follows from the fact that if \( b \in (−\infty,t]\cap S \), then \( t \geq b \). Since \( b > a \), we must have \( t > a \), so that \( a \in (−\infty,t]\cap S \) as well. Since \( a \) and \( b \) are arbitrary, we see that no two-point subset of \( \mathbb{R} \) can be shattered by \( \mathcal{C} \).

2.2. Closed intervals. Again, let \( Z = \mathbb{R} \) and take \( \mathcal{C} \) to be the class of all intervals of the form \([s,t]\) for all \( s, t \in \mathbb{R} \). Then \( V(\mathcal{C}) = 2 \). To see this, we will show that (1) any two point set \( S = \{a,b\} \) can be shattered by \( \mathcal{C} \) and that (2) no three-point set \( S = \{a,b,c\} \) can be shattered by \( \mathcal{C} \).

For (1), let \( S = \{a,b\} \) and suppose, without loss of generality, that \( a < b \). Choose four points \( t_1, t_2, t_3, t_4 \in \mathbb{R} \) such that \( t_1 < t_2 < a < t_3 < b < t_4 \). There are four subsets of \( S \): \( \emptyset \), \( \{a\} \), \( \{b\} \), and \( \{a,b\} = S \). Then

\[
[t_1, t_2] \cap S = \emptyset, \quad [t_2, t_3] \cap S = \{a\}, \quad [t_3, t_4] \cap S = \{b\}, \quad [t_1, t_4] \cap S = S.
\]

Hence, \( S \) is shattered by \( \mathcal{C} \). This holds for every two-point set in \( \mathbb{R} \), which proves (1). To prove (2), let \( S = \{a,b,c\} \) be an arbitrary three-point set with \( a < b < c \). Then the intersection of any \( [t_1, t_2] \in \mathcal{C} \) with \( S \) containing \( a \) and \( c \) must necessarily contain \( b \) as well. This shows that no three-point set can be shattered by \( \mathcal{C} \), so by Lemma 1 we conclude that \( V(\mathcal{C}) = 2 \).

2.3. Closed halfspaces. Let \( Z = \mathbb{R}^2 \), and let \( \mathcal{C} \) consist of all closed halfspaces, i.e., sets of the form

\[
\{ z = (z_1, z_2) \in \mathbb{R}^2 : w_1 z_1 + w_2 z_2 \geq b \}
\]

for all choices of \( w_1, w_2, b \in \mathbb{R} \) such that \( (w_1, w_2) \neq (0,0) \). Then \( V(\mathcal{C}) = 3 \).

To see that \( S_3(\mathcal{C}) = 2^3 = 8 \), it suffices to consider any set \( S = \{z_1, z_2, z_3\} \) of three non-collinear points. Then it is not hard to see that for any \( S' \subseteq S \) it is possible to choose a closed halfspace \( C \in \mathcal{C} \) that would contain \( S' \), but not \( S \). To see that \( S_4(\mathcal{C}) < 2^4 \), we must look at all four-point sets \( S = \{z_1, z_2, z_3, z_4\} \). There are two cases to consider:

1. One point in \( S \) lies in the convex hull of the other three. Without loss of generality, let’s suppose that \( z_1 \in \text{conv}(S') \) with \( S' = \{z_2, z_3, z_4\} \). Then there is no \( C \in \mathcal{C} \) such that \( C \cap S = S' \). The reason for this is that every \( C \in \mathcal{C} \) is a convex set. Hence, if \( S' \subset C \), then any point in \( \text{conv}(S') \) is contained in \( C \) as well.
No point in $S$ is in the convex hull of the remaining points. This case, when $S$ is an affinely independent set, is shown in Figure 1. Let us partition $S$ into two disjoint subsets, $S_1$ and $S_2$, each consisting of “opposite” points. In the figure, $S_1 = \{z_1, z_3\}$ and $S_2 = \{z_2, z_4\}$. Then it is easy to see that there is no halfspace $C$ whose boundary could separate $S_1$ from its complement $S_2$. This is, in fact, the famous “XOR counterexample” of Minsky and Papert [MP69], which has demonstrated the impossibility of universal concept learning by one-layer perceptrons.

Since any four-point set in $\mathbb{R}^2$ falls under one of these two cases, we have shown that no such set can be shattered by $C$. Hence, $V(C) = 3$.

More generally, if $Z = \mathbb{R}^d$ and $C$ is the class of all closed halfspaces

\[
\left\{ z \in \mathbb{R}^d : \sum_{j=1}^{d} w_j z_j \geq b \right\}
\]

for all $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ such that at least one of the $w_j$’s is nonzero and all $b \in \mathbb{R}$, then $V(C) = d + 1$ [WD81]; we will see a proof of this fact shortly.

2.4. Axis-parallel rectangles. Let $Z = \mathbb{R}^2$, and let $C$ consist of all “axis-parallel” rectangles, i.e., sets of the form $C = [a_1, b_1] \times [a_2, b_2]$ for all $a_1, b_1, a_2, b_2 \in \mathbb{R}$. Then $V(C) = 4$.

First we exhibit a four-point set $S = \{z_1, z_2, z_3, z_4\}$ that is shattered by $C$. It suffices to take $z_1 = (-1, -1)$, $z_2 = (-1, 1)$, $z_3 = (1, -1)$, $z_4 = (1, 1)$. To show that no five-point set is shattered by $C$, consider an arbitrary $S = \{z_1, z_2, z_3, z_4, z_5\}$. Of these, pick any one point with the smallest first coordinate and any one point with the largest first coordinate, and likewise for the second coordinate (refer to Figure 2), for a total of at most four. Let $S'$ denote the set consisting of these points; in Figure 2, $S' = \{z_1, z_2, z_3, z_4\}$. Then it is easy to see that any $C \in C$ that contains the points in $S'$ must contain all the points in $S \setminus S'$ as well. Hence, no five-point set in $\mathbb{R}^2$ can be shattered by $C$, so $V(C) = 5$.

The same argument also works for axis-parallel rectangles in $\mathbb{R}^d$, i.e., all sets of the form $C = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$, leading to the conclusion that the VC dimension of the set of all axis-parallel rectangles in $\mathbb{R}^d$ is equal to $2d$.

2.5. Sets determined by finite-dimensional function spaces. The following result is due to Dudley [Dud78]. Let $Z$ be arbitrary, and let $\mathcal{G}$ be an $m$-dimensional linear space of functions
Figure 2. Impossibility of shattering a five-point set by axis-parallel rectangles.

\( g : \mathbb{Z} \to \mathbb{R} \), which means that each \( g \in \mathcal{G} \) has a unique representation of the form

\[
g = \sum_{j=1}^{m} c_j \psi_j,
\]

where \( \psi_1, \ldots, \psi_m : \mathbb{Z} \to \mathbb{R} \) form a fixed linearly independent set and \( c_1, \ldots, c_m \) are real coefficients. Consider the class

\[
\mathcal{C} = \left\{ \{ z \in \mathbb{Z} : g(z) \geq 0 \} : g \in \mathcal{G} \right\}.
\]

Then \( V(\mathcal{C}) \leq m \).

To prove this, we need to show that no set of \( m + 1 \) points in \( \mathbb{Z} \) can be shattered by \( \mathcal{C} \). To that end, let us fix \( m + 1 \) arbitrary points \( z_1, \ldots, z_{m+1} \in \mathbb{Z} \) and consider the mapping \( L : \mathcal{G} \to \mathbb{R}^{m+1} \) defined by

\[
L(g) \triangleq (g(z_1), \ldots, g(z_{m+1})).
\]

It is easy to see that because \( \mathcal{G} \) is a linear space, \( L \) is a linear mapping, i.e., for any \( g_1, g_2 \in \mathcal{G} \) and any \( c_1, c_2 \in \mathbb{R} \) we have \( L(c_1 g_1 + c_2 g_2) = c_1 L(g_1) + c_2 L(g_2) \). Since \( \dim \mathcal{G} = m \), the image of \( \mathcal{G} \) under \( L \), i.e., the set

\[
L(\mathcal{G}) = \left\{ (g(z_1), \ldots, g(z_{m+1})) \in \mathbb{R}^{m+1} : g \in \mathcal{G} \right\},
\]

is a linear subspace of \( \mathbb{R}^{m+1} \) of dimension at most \( m \). This means that there exists some nonzero vector \( v = (v_1, \ldots, v_{m+1}) \in \mathbb{R}^{m+1} \) orthogonal to \( L(\mathcal{G}) \), i.e., for every \( g \in \mathcal{G} \)

\[
v_1 g(z_1) + \ldots + v_{m+1} g(z_{m+1}) = 0.
\]

Without loss of generality, we may assume that at least one component of \( v \) is strictly negative (otherwise we can take \( -v \) instead of \( v \) and still get (3)). Hence, we can rearrange the equality in (3) as

\[
\sum_{i : v_i \geq 0} v_i g(z_i) = - \sum_{i : v_i < 0} v_i g(z_i), \quad \forall g \in \mathcal{G}.
\]

Now let us suppose that \( \mathbb{S}_{m+1}(\mathcal{C}) = 2^{m+1} \) and derive a contradiction. Consider a binary \((m + 1)\)-tuple \( b = (b_1, \ldots, b_{m+1}) \in \{0, 1\}^{m+1} \), where \( b_j = 1 \) if and only if \( v_j \geq 0 \), and 0 otherwise. Since we assumed that \( \mathbb{S}_{m+1}(\mathcal{C}) = 2^{m+1} \), there exists some \( g \in \mathcal{G} \) such that

\[
\left( 1_{\{ g(z_1) \geq 0 \}}, \ldots, 1_{\{ g(z_{m+1}) \geq 0 \}} \right) = b.
\]

By our definition of \( b \), this means that the left-hand side of (4) is nonnegative, while the right-hand side is negative, which is a contradiction. Hence, \( \mathbb{S}_{m+1}(\mathcal{C}) < 2^{m+1} \), so \( V(\mathcal{C}) \leq m \).

This result can be used to bound the VC dimension of many classes of sets:
Let $C$ be the class of all closed halfspaces in $\mathbb{R}^d$. Then any $C \in C$ can be represented in the form $C = \{ z : g(z) \geq 0 \}$ for $g(z) = \langle w, z \rangle - b$ with some nonzero $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The set $G$ of all such affine functions on $\mathbb{R}^d$ is a linear space of dimension $d + 1$, so by the above result we have $V(C) \leq d + 1$. In fact, we know that this holds with equality [WD81].

This can also be seen from the following result, due to Cover [Cov65]: Let $G$ be the linear space of functions spanned by functions $\psi_1, \ldots, \psi_m$, and let $\{ z_1, \ldots, z_n \} \subset \mathbb{Z}$ be such that the vectors $\Psi(z_i) = (\psi_1(z_i), \ldots, \psi_m(z_i)), 1 \leq i \leq n$, form a linearly independent set. Then for the class of sets $C = \{ z : g(z) \geq 0 \} : z \in \mathbb{Z}$ we have

$$|C \cap \{ z_1, \ldots, z_n \} : C \in C| = \sum_{i=0}^{m-1} \binom{n-1}{i}.$$

The conditions needed for Cover’s result are seen to hold for indicators of halfspaces, so letting $n = m = d + 1$ we see that $S_d(C) = \sum_{i=0}^{d} \binom{d}{i} = 2^d$. Hence, $V(C) = d + 1$.

Let $C$ be the class of all closed balls in $\mathbb{R}^d$, i.e., sets of the form

$$C = \left\{ z \in \mathbb{R}^d : ||z - x||^2 \leq r^2 \right\}$$

where $x \in \mathbb{R}^d$ is the center of $C$ and $r \in \mathbb{R}^+$ is its radius. Then we can write $C = \{ z : g(z) \geq 0 \}$, where

$$g(z) = r^2 - ||z - x||^2 = r^2 - \sum_{j=1}^{d} |z_j - x_j|^2.$$

Expanding the second expression for $g$ in (5), we get

$$g(z) = r^2 - \sum_{j=1}^{d} x_j^2 + 2 \sum_{j=1}^{d} x_j z_j - \sum_{j=1}^{d} z_j^2,$$

which can be written in the form $g(z) = \sum_{k=1}^{d+2} c_k \psi_k(z)$, where $\psi_1(z) = 1, \psi_k(z) = z_k$ for $k = 2, \ldots, d + 1$, and $\psi_{d+2} = \sum_{j=1}^{d} z_j^2$. It can be shown that the functions $\{ \psi_k \}_{k=1}^{d+2}$ are linearly independent. Hence, $V(C) \leq d + 2$. This bound, however, is not tight; as shown by Dudley [Dud79], the class of closed balls in $\mathbb{R}^d$ has VC dimension $d + 1$.


The importance of VC classes in learning theory arises from the fact that, as $n$ tends to infinity, the fraction of subsets of any $\{ z_1, \ldots, z_n \} \subset \mathbb{Z}$ that are shattered by a given VC class $C$ tends to zero. We will prove this fact in this section by deriving a sharp bound on the shatter coefficients $S_n(C)$ of a VC class $C$. This bound have been (re)discovered at least three times, first in a weak form by Vapnik and Chervonenkis [VC71] in 1971, then independently and in different contexts by Sauer [Sau72] and Shelah [She72] in 1972. In strict accordance with Stigler’s law of eponymy\(^1\), it is known in the statistical learning literature as the Sauer–Shelah lemma.

Before we state and prove this result, we will collect some preliminaries and set up some notation. Given integers $n, d \geq 1$, let

$$\phi(n, d) \triangleq \begin{cases} 
\sum_{i=0}^{d} \binom{n}{i}, & \text{if } n > d \\
2^n, & \text{if } n \leq d 
\end{cases}$$

\(^{1}\)“No scientific discovery is named after its original discoverer” (http://en.wikipedia.org/wiki/Stigler’s_law_of_epony)
If we adopt the convention that \( \binom{i}{n} = 0 \) for \( i > n \), we can write

\[
\phi(n, d) = \sum_{i=0}^{d} \binom{n}{i}
\]

for all \( n, d \geq 1 \). We will find the following recursive relation useful:

**Lemma 2.**

\[
\phi(n, d) = \phi(n - 1, d) + \phi(n - 1, d - 1).
\]

**Proof.** We have

\[
\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i)!} + \frac{(n-1)!}{i!(n-i-1)!}.
\]

Multiplying both sides by \( i!(n-i)! \), we obtain

\[
i!(n-i)! \left[ \binom{n-1}{i-1} + \binom{n-1}{i} \right] = i(n-1)! + (n-i)(n-1)! = n!
\]

Hence,

\[
\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{n!}{i!(n-i)!} = \binom{n}{i}.
\]

Using the definition of \( \phi(n, d) \), as well as (6), we get

\[
\phi(n, d) = \sum_{i=0}^{d} \binom{n}{i} = 1 + \sum_{i=1}^{d} \binom{n}{i} = 1 + \sum_{i=1}^{d} \binom{n-1}{i-1} + \sum_{i=1}^{d} \binom{n-1}{i} = \phi(n-1, d) + \phi(n-1, d-1)
\]

and the lemma is proved.

Now for the actual result:

**Theorem 1** (Sauer–Shelah lemma). Let \( C \) be a class of subsets of some space \( Z \) with \( V(C) = d < \infty \). Then for all \( n \),

\[
S_n(C) \leq \phi(n, d).
\]

**Proof.** There are several different proofs in the literature; we will use an inductive argument following Blumer et al. [BEHW89].

We can assume, without loss of generality, that \( n > d \), for otherwise \( S_n(C) = 2^n = \phi(n, d) \). For an arbitrary finite set \( S \subset Z \), let

\[
S(S, C) \triangleq \{ S \cap C : C \in C \},
\]

where, as before, we count only the distinct sets of the form \( S \cap C \). By definition, \( S_n(C) = \sup_{|S|=n} S(S, C) \). Thus, it suffices the prove the following: For any \( S \subset Z \) with \( |S| = n > d \),

\[
S(S, C) \leq \phi(n, d).
\]

For the purpose of computing \( S(S, C) \), any two \( C_1, C_2 \in C \) such that \( S \cap C_1 = S \cap C_2 \) are deemed equivalent. Hence, let

\[
\mathcal{A} \triangleq \{ A \subseteq S : A = S \cap C \text{ for some } C \in C \}.
\]

Then we may write

\[
S(S, C) = |\{ S \cap C : C \in C \}| = |\{ A \subseteq S : A = S \cap C \text{ for some } C \in C \}| = |\mathcal{A}|.
\]

Moreover, it is easy to see that \( V(\mathcal{A}) \leq V(C) = d \).

Thus, the desired result is equivalent to saying that if \( \mathcal{A} \) is a collection of subsets of an \( n \)-element set \( S \) (which we may, without loss of generality, take to be \( [n] \triangleq \{1, \ldots, n\} \)) with \( V(\mathcal{A}) \leq d < n \),
then $|A| \leq \phi(n, d)$. We will prove this statement by “double induction” on $n$ and $d$. First of all, the statement (7) holds for all $n \geq 1$ and $d = 0$. Indeed, if $V(A) = 0$, then $|A| = 1 \leq 2^n$. Now assume that (7) holds for all $n$ and all $A$ with $V(A) \leq d - 1$, and for all integers up to $n - 1$ and all $A$ with $V(A) \leq d$. Now let $S = [n]$, and let $A$ be a collection of subsets of $[n]$ with $V(A) = d < n$. We will show that $|A| \leq \phi(n, d)$.

To prove this claim, let us choose an arbitrary $i \in S$ and define

$$A\{i\} \triangleq \{A\{i\} : A \in A\}$$

$$A_i \triangleq \{A \in A : i \not\in A, A \cup \{i\} \in A\}$$

Observe that both $A\{i\}$ and $A_i$ are classes of subsets of $S \setminus \{i\}$. Moreover, since $A$ and $A \cup \{i\}$ map to the same element of $A\{i\}$, while $|A_i|$ is the number of pairs of sets in $A$ that map into the same set in $A\{i\}$, we have

(8) $$|A| = |A\{i\}| + |A_i|.$$  

Since $A\{i\} \subseteq A$, we have $V(A\{i\}) \leq V(A) \leq d$. Also, every set in $A\{i\}$ is a subset of $S \setminus \{i\}$, which has cardinality $n - 1$. Therefore, by the inductive hypothesis $|A\{i\}| \leq \phi(n - 1, d)$. Next, we show that $V(A_i) \leq d - 1$. Suppose, to the contrary, that $V(A_i) = d$. Then there must exist some $T \subseteq S \setminus \{i\}$ with $|T| = d$ that is shattered by $A_i$. But then $T \cup \{i\}$ is shattered by $A$. To see this, given any $T' \subseteq T$ choose some $A \in A_i$ such that $T \cap A = T'$ (this is possible since $T$ is shattered by $A_i$). But then $A \cup \{i\} \in A$ (by definition of $A_i$), and

$$(T \cup \{i\}) \cap (A \cup \{i\}) = (T \cap A) \cup \{i\} = T' \cup \{i\}.$$  

Since this is possible for an arbitrary $T' \subseteq T$, we conclude that $T \cup \{i\}$ is shattered by $A$. Now, since $T \subseteq S \setminus \{i\}$, we must have $i \not\in T$, so $|T \cup \{i\}| = |T| + 1 = d + 1$, which means that there exists a $(d + 1)$-element subset of $S = [n]$ that is shattered by $A$. But this contradicts our assumption that $V(A) \leq d$. Hence, $V(A_i) \leq d - 1$. Since $A_i$ is a collection of subsets of $S \setminus \{i\}$, we must have $|A_i| \leq \phi(n - 1, d - 1)$ by the inductive hypothesis. Hence, from (8) and from Lemma 2 we have

$$|A| = |A\{i\}| + |A_i| \leq \phi(n - 1, d) + \phi(n - 1, d - 1) = \phi(n, d).$$

This completes the induction argument and proves (7).

\[ \square \]

**Corollary 1.** If $C$ is a collection of sets with $V(C) \leq d < \infty$, then

$$\mathbb{S}_n(C) \leq (n + 1)^d.$$  

Moreover, if $n \geq d$, then

$$\mathbb{S}_n(C) \leq \left(\frac{en}{d}\right)^d,$$

where $e$ is the base of the natural logarithm.

**Proof.** For the first bound, write

$$\phi(n, d) = \sum_{i=0}^{d} \binom{n}{i} = \sum_{i=1}^{d} \frac{n!}{i!(n-i)!} \leq \sum_{i=1}^{d} \frac{n^i}{i!} \leq \sum_{i=1}^{d} \frac{n^i d!}{i!(d-i)!} = \sum_{i=0}^{d} \binom{d}{i} \binom{n}{i} = (1 + \frac{d}{n})^n \leq e^d,$$

where the last step uses the binomial theorem. On the other hand, if $d/n \leq 1$, then

$$\left(\frac{d}{n}\right)^d \phi(n, d) = \left(\frac{d}{n}\right)^d \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=1}^{d} \left(\frac{d}{n}\right)^i \frac{n}{i} \leq \sum_{i=1}^{d} \left(\frac{d}{n}\right)^i \binom{n}{i} = \left(1 + \frac{d}{n}\right)^n \leq e^d,$$

where we again used the binomial theorem. Dividing both sides by $(d/n)^d$, we get the second bound. \[ \square \]
Let $C$ be a VC class of subsets of some space $Z$. From the above corollary we see that
\[
\limsup_{n \to \infty} \frac{S_n(C)}{2^n} \leq \lim_{n \to \infty} \frac{(n+1)^{V(C)}}{2^n} = 0.
\]
In other words, as $n$ becomes large, the fraction of subsets of an arbitrary $n$-element set $\{z_1, \ldots, z_n\} \subset Z$ that are shattered by $C$ becomes negligible. Moreover, combining the bounds of the corollary with the Finite Class Lemma for Rademacher averages, we get the following:

**Theorem 2.** Let $F$ be a VC class of binary-valued functions $f : Z \to \{0, 1\}$ on some space $Z$. Let $Z^n$ be an i.i.d. sample of size $n$ drawn according to an arbitrary probability distribution $P \in \mathcal{P}(Z)$. Then
\[
\mathbb{E} R_n(F(Z^n)) \leq 2\sqrt{V(F) \log(n+1)}.
\]
A more refined chaining technique [Dud78] can be used to remove the logarithm in the above bound:

**Theorem 3.** There exists an absolute constant $C > 0$, such that under the conditions of the preceding theorem
\[
\mathbb{E} R_n(F(Z^n)) \leq C\sqrt{\frac{V(F)}{n}}.
\]

**References**


