EECS 598: Statistical Learning Theory, Winter 2014 **Topic 5**

Vapnik-Chevronenkis Theory

Lecturer: Clayton Scott *Scribe: Srinagesh Sharma, Scott Reed, Petter Nilsson*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

Let's say we are given training data $\{(X_i, Y_i)\}_{i=1}^n$ which are drawn from $\mathcal{X} \times \mathcal{Y}$ independently from the distribution P_{XY}, and a set of classifiers $\mathcal{H}$. In the previous lecture, we saw that performance guarantees for empirical risk minimization over $\mathcal{H}$ follow from uniform deviation bounds of the form

$$\Pr\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - R(h)\right| \geq \epsilon\right) \leq \delta$$

where $\widehat{R}_n(h) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{h(X_i)\neq Y_i\}}$ is the empirical risk and $R(h)$ is the true risk. We also established such a bound for finite $\mathcal{H}$. In these notes we turn our attention to the setting where $\mathcal{H}$ is infinite, and possibly uncountable. This will lead us to an interesting notion of the capacity of $\mathcal{H}$ known as the Vapnik-Chervonenkis dimension.

2 VC Theorem

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. For $x_1, x_2, \ldots, x_n \in \mathcal{X}$ denote

$$N_{\mathcal{H}}(x_1, \ldots, x_n) := |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|$$

Clearly $N_{\mathcal{H}}(x_1, \ldots, x_n) \leq 2^n$. The n^{th} *shatter coefficient* is defined as

$$S_{\mathcal{H}}(n) =: \max_{x_1,\ldots,x_n\in\mathcal{X}} N_{\mathcal{H}}(x_1, \ldots, x_n)$$

If $S_{\mathcal{H}}(n) = 2^n$, then $\exists\ x_1, \ldots, x_n$ such that

$$N_{\mathcal{H}}(x_1, \ldots, x_n) = 2^n$$

and we say that $\mathcal{H}$ *shatters* $x_1, \ldots, x_n$.

Note. The shatter coefficient is sometimes called the *growth function* in the literature.

The *VC dimension* of $\mathcal{H}$ is defined as

$$V_{\mathcal{H}} := \max\{n \mid S_{\mathcal{H}}(n) = 2^n\}.$$

If $S_{\mathcal{H}}(n) = 2^n\ \forall n$ then $V_{\mathcal{H}} := \infty$.

Remark. To show $V_{\mathcal{H}} = V$ we must show that there exists at least one set of points $x_1, \ldots, x_n$ that can be shattered by $\mathcal{H}$, and that *no* set of $n+1$ points can be shattered by $\mathcal{H}$.

The VC dimension and the shatter coefficient relate to the following uniform deviation bound.

Theorem 1. *For any $n \geq 1$ and $\epsilon > 0$*

$$\Pr\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - R(h)\right| \geq \epsilon\right) \leq 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/32}. \tag{1}$$

where the probability is with respect to the draw of the training data.

We will show below that $S_{\mathcal{H}}(n) \leq (n+1)^{V_{\mathcal{H}}}$. Therefore if $V_{\mathcal{H}}$ is finite then the right hand side of equation (1) is dominated by the exponential and will go to zero as $n \to \infty$. Similar to the case $|\mathcal{H}| < \infty$ we also have a performance guarantee for ERM when $V_{\mathcal{H}} < \infty$.

Corollary 1. *If $\widehat{h}_n$ is an empirical risk minimizer (ERM) over $\mathcal{H}$ then*

$$\Pr\left(R(\widehat{h}_n) - R^*_{\mathcal{H}} \geq \epsilon\right) \leq 8S_{\mathcal{H}}(n)e^{-n\epsilon^2/128},$$

where the probability is with respect to the draw of the training data. Equivalently, with probability greater than $1-\delta$

$$\begin{aligned} R(\widehat{h}_n) &\leq R^*_{\mathcal{H}} + \sqrt{\frac{128[\log(\frac{8S_{\mathcal{H}}(n)}{\delta})]}{n}} \\ &\leq R^*_{\mathcal{H}} + \sqrt{\frac{128[V_{\mathcal{H}}\log(n+1) + \log(\frac{8}{\delta})]}{n}}. \end{aligned}$$

Proof. Follows from Theorem 1 using an argument like that when $|\mathcal{H}| < \infty$. □

Corollary 2. *If $V_{\mathcal{H}} < \infty$ then $\mathcal{H}$ is uniformly learnable by ERM.*

The version of the VC inequality given in Theorem 1 is proved in [1]. That reference also contains a broader discussion of VC theory than is presented here. We will prove the VC inequality later (with perhaps different constants) after discussing Rademacher complexity.

2.1 VC Classes

A *VC class* is a set of classifiers $\mathcal{H}$ with $V_{\mathcal{H}} < \infty$. We will now consider some examples of $\mathcal{H}$ for which $V_{\mathcal{H}}$ can be established or at least bounded.

Example. Consider the set of classifiers

$$\mathcal{H} = \left\{\mathbf{1}_{\{x\in R\}}\middle| R = \prod_{i=1}^{d}[a_i, b_i],\ \ a_i < b_i\right\}. \qquad \ast$$

Let $d = 1$. Given one point, we can always assign it a one or a zero. Therefore $V_{\mathcal{H}} > 1$. For two points there are four possible assignments and they can be shattered using $\mathcal{H}$. However, given 3 points, the following assignment cannot be realized by any $h \in \mathcal{H}$. Therefore $N_{\mathcal{H}}(x_1, \ldots, x_n) < 8$, and so $V_{\mathcal{H}} = 2$.

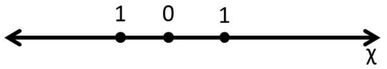

Figure 1: This classification does not belong to $N_{\mathcal{H}}(x_1, \ldots, x_n)$ when d=1

For $d = 2$, the four points in Fig. 2 can be shattered by $\mathcal{H}$. Therefore $V_{\mathcal{H}}$ is at least 4. For $n = 5$, there is a maximum and minimum point in each dimension. Consider a set of ≤ 4 points achieving these

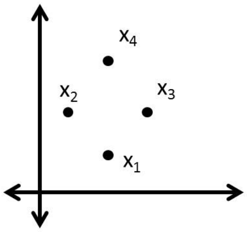

Figure 2: Four point shattered when $d = 2$

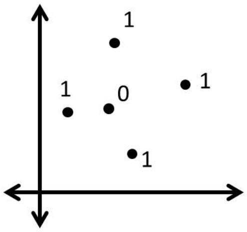

Figure 3: No five points can be shattered when $d = 2$

extrema. Now there exists at least 1 point not among these four. If it is in the interior of the bounding rectangle, it cannot be labeled 0 while the others are labeled 1 (see Figure 3). If it is on the boundary of the bounding rectangle, it cannot be assigned a different label from the other point achieving the same face of the boundary. Therefore $V_{\mathcal{H}} = 4$.

In general, for dimension d, $\mathcal{H}$ can shatter the $2d$ points

$$\begin{gathered}(\pm 1, 0, \dots, 0)\\(0, \pm 1, \dots, 0)\\\vdots\\(0, 0, \dots, \pm 1)\end{gathered}$$

and it cannot shatter any $2d + 1$ points by an argument like that when $d = 2$. Therefore $V_{\mathcal{H}} = 2d$.

Example. Let $\mathcal{X} = \mathbb{R}^2$. Consider the set of classifiers ✱

$$\mathcal{H} = \left\{ \mathbf{1}_{\{x \in C\}} \,\middle|\, \text{C is convex in } \mathbb{R}^2 \right\}.$$

For any n, consider n points on a circle. Regardless of how these points are labeled, there always exists a polygon configuration that realizes those labels (see Fig. 4). Therefore $V_{\mathcal{H}} = \infty$.

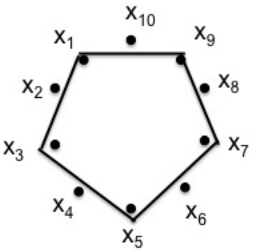

Figure 4: Classifiers based on convex sets can shatter any number of points lying on a circle. Here $n = 10$.

Example. Consider the case where $|\mathcal{H}| < \infty$. Then we have ✱

$$\begin{aligned} N_{\mathcal{H}}(x_1, \dots, x_n) &= |\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}| \leq |\mathcal{H}| \\ &\Longrightarrow S_{\mathcal{H}}(n) \leq |\mathcal{H}| \\ &\Longrightarrow V_{\mathcal{H}} \leq \log_2(|\mathcal{H}|). \end{aligned}$$

As a sanity check, this results in the bound

$$\Pr\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h)-R(h)\right|\geq\epsilon\right)\leq 8\,|\mathcal{H}|\,e^{-n\epsilon^2/32}$$

which is basically the same as what was derived previously except for larger constants.

The following result lets us bound the VC dimension of a broad family of classes.

Lemma 1. *Let $\mathcal{F}$ be an m-dimensional vector space of real-valued functions. Then,* *

$$\mathcal{H}=\left\{\mathbf{1}_{\{f(x)\geq 0\}}\middle|f\in\mathcal{F}\right\}$$

has $V_{\mathcal{H}}\leq m$.

Proof. Suppose $\mathcal{H}$ shatters $m+1$ points, say $x_1,\ldots,x_{m+1}$. Define the linear mapping $L:\mathcal{F}\to\mathbb{R}^{m+1}$,

$$L(f)=(f(x_1),\ldots,f(x_{m+1}))^T.$$

Since $\dim(\mathcal{F})=m$ we have $\dim(L(\mathcal{F}))\leq m$ where $L(\mathcal{F})$ denotes the image of $\mathcal{F}$. By the projection theorem,

$$\mathbb{R}^{m+1}=L(\mathcal{F})\oplus L(\mathcal{F})^{\perp}$$

where $\oplus$ denotes the direct sum. Therefore

$$\dim(L(\mathcal{F})^{\perp})\geq 1$$

and so there exits $\gamma\neq 0,\ \gamma\in\mathbb{R}^{m+1}$ such that

$$\gamma^T L(f)=0\quad\forall f\in\mathcal{F}.$$

Thus, $\forall f\in\mathcal{F}$

$$\sum_{i=1}^{m+1}\gamma_i f(x_i)=0$$

or equivalently,

$$\sum_{i:\gamma_i\geq 0}\gamma_i f(x_i)=\sum_{i:\gamma_i<0}-\gamma_i f(x_i).$$

We will assume that at least one $\gamma_i<0$. If not, replace γ by $-\gamma$. Since $\mathcal{H}$ shatters $x_1,\ldots,x_{m+1}$, let $h=1_{\{f(x)\geq 0\}}$ be such that

$$h(x_i)=1\iff\gamma_i\geq 0.$$

For the corresponding f

$$f(x_i)\geq 0\iff\gamma_i\geq 0.$$

This implies that $\sum_{i:\gamma_i\geq 0}\gamma_i f(x_i)\geq 0$ and $\sum_{i:\gamma_i<0}-\gamma_i f(x_i)<0$ which is a contradiction. Therefore, $V_{\mathcal{H}}\leq m$. □

Let's apply this result to the class of linear classifiers. *

Example. Let $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{F}=\left\{f\,\middle|\,f(x)=w^Tx+b,w\in\mathbb{R}^d,b\in\mathbb{R}\right\}$. Then $\mathcal{H}$ is the set of all linear classifiers,

$$\mathcal{H}=\left\{\mathbf{1}_{\{w^Tx+b\geq 0\}}\middle|w\in\mathbb{R}^d,b\in\mathbb{R}\right\}.$$

Since $\dim(\mathcal{F})=d+1$ we deduce from the lemma that $V_{\mathcal{H}}\leq d+1$. In fact, this bound is achieved so that $V_{\mathcal{H}}=d+1$. For $d=2$, $\mathcal{H}$ shatters the vertices of any nondegenerate triangle, for $d=3$, $\mathcal{H}$ shatters the vertices of a tetrahedron, and for general d, $\mathcal{H}$ shatters the zero vector along with the standard basis in $\mathbb{R}^d$.

3 Sauer's Lemma

This is a bound on the shatter coefficient that was proved independently by Vapnik and Chervonenkis (1971), Sauer (1972) and Shelah (1972).

Theorem 2. *Let $V = V_{\mathcal{H}} < \infty$. For all $n \geq 1$,*

$$S_{\mathcal{H}}(n) \leq \sum_{i=0}^{V} \binom{n}{i}.$$

Before proving this theorem we consider several corollaries:

Corollary 3. *If $V < \infty$, then $\forall n \geq 1$,*

$$S_{\mathcal{H}}(n) \leq (n+1)^V.$$

Proof. By the binomial theorem,

$$\begin{aligned}(n+1)^V &= \sum_{i=0}^{V} n^i \binom{V}{i} = \sum_{i=0}^{V} n^i \frac{V!}{(V-i)!i!} \\ &\geq \sum_{i=0}^{V} \frac{n^i}{i!} \geq \sum_{i=0}^{V} \frac{n!}{(n-i)!i!} = \sum_{i=0}^{V} \binom{n}{i}.\end{aligned}$$

□

Corollary 4. $\forall n \geq V$,

$$S_{\mathcal{H}}(n) \leq \left(\frac{ne}{V}\right)^V.$$

Proof. If $\frac{V}{n} \leq 1$ then

$$\begin{aligned}\left(\frac{V}{n}\right)^V \sum_{i=0}^{V} \binom{n}{i} &\leq \sum_{i=0}^{V} \left(\frac{V}{n}\right)^i \binom{n}{i} \leq \sum_{i=0}^{n} \left(\frac{V}{n}\right)^i \binom{n}{i} \\ &= \left(1 + \frac{V}{n}\right)^n \leq e^V\end{aligned}$$

Therefore

$$\sum_{i=0}^{V} \binom{n}{i} \leq \left(\frac{ne}{V}\right)^V.$$

□

Corollary 5. *If $V > 2$, then $\forall n \geq V$,*

$$S_{\mathcal{H}}(n) \leq n^V.$$

Proof. If $V > 2$, then $\frac{e}{V} < 1$, so the statement holds by Corollary 4. □

Proof of Sauer's Lemma. For $n \leq V$,

$$\sum_{i=0}^{V} \binom{n}{i} \geq \sum_{i=0}^{n} \binom{n}{i} = (1+1)^n = 2^n = S_{\mathcal{H}}(n)$$

Assume $n > V$. We will show $\forall\, x_1, x_2, ..., x_n$

$$N_{\mathcal{H}}(x_1, x_2, ..., x_n) \leq \sum_{i=0}^{V} \binom{n}{i}.$$

So fix $x_1, x_2, ..., x_n \in \mathcal{X}$. Define the matrix B_0 to have rows consisting of all possible values of $(h(x_1), ..., h(x_n))$. B_0 has dimension $N_{\mathcal{H}} \times n$. Define $B_i, i = 1, ..., n$ by the following iterative procedure:

- For each row in column i of B_{i-1} replace each 1 by a 0 unless it produces another row of B_{i-1}.

Example. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{1_{\{x \in [a,b]\}} | a, b \in \mathbb{R}, a < b\}$. Then $V = 2$. Let's take $n = 4$. Then

$$B_0 = \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array}\right] \end{array}, \quad B_4 = \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{array}\right] \end{array}.$$

The following three properties hold in general for this construction. For the third property, we say a binary matrix B with n columns *shatters* an index set $\mathcal{I} \subseteq \{1, ..., n\}, |\mathcal{I}| = m$, if $B|_{\mathcal{I}}$ contains all elements of $\{0, 1\}^m$.

1. $|B_0| = |B_n|$; rows remain distinct.
2. In B_n, replacing any 1 with a 0 produces another row of B_n.
3. If B_n shatters an index set, then so does B_0.

The first two proporties follow from the construction. To show the third property, it suffices to show that if B_i shatters $\mathcal{I}$, then so does B_{i-1}. Without loss of generality take $i = 1$. Suppose B_1 shatters $\mathcal{I} \subseteq \{1, ..., n\}$, $|\mathcal{I}| = m$. If $1 \notin \mathcal{I}$, then clearly B_0 shatters $\mathcal{I}$. Suppose $1 \in \mathcal{I}$. Let $\mathcal{I} = \{i_1, ..., i_m\}$, $i_1 = 1$. Let $(b_{i_1}, b_{i_2}, ..., b_{i_m}) \in \{0, 1\}^m$. Since B_1 shatters $\mathcal{I}$, we know $(1, b_{i_2}, ...b_{i_m})$ is a row of $B_1|_{\mathcal{I}}$. Then both $(0, b_{i2}, ..., b_{im})$ and $(1, b_{i2}, ..., b_{im})$ are rows of $B_0|_{\mathcal{I}}$ by definition of B_1. Thus B_0 shatters $\mathcal{I}$.

By 2) and 3), B_n cannot have $> V$ 1s in any row. If it did, B_n would shatter those points, and hence so would B_0. By 1), $N_{\mathcal{H}}(x_1, ..., x_n) = |B_0| = |B_n| \leq \sum_{i=0}^{V} \binom{n}{i}$. The last term is a bound on the number ways the 1s can occur in the rows. □

4 VC Theory for Sets

Let $\mathcal{G} \subset 2^{\mathcal{X}}$. We can define

$$N_{\mathcal{G}}(x_1, ..., x_n) = |\{G \cap \{x_1, ..., x_n\} : G \in \mathcal{G}\}|$$
$$S_{\mathcal{G}}(n) = \max_{x_1, ..., x_n} N_{\mathcal{G}}(x_1, ..., x_n)$$
$$V_{\mathcal{G}} = \max\{n : S_{\mathcal{G}}(n) = 2^n\}$$

in analogy to our definitions for classifiers. Indeed, sets and binary classifiers are equivalent via

$$\begin{aligned} G &\mapsto h_G(x) = 1_{\{x \in G\}} \\ h &\mapsto G_h = \{x : h(x) = 1\}. \end{aligned}$$

This gives a VC theorem for sets.

Corollary 6. *If $X_1, ..., X_n \overset{iid}{\sim} Q$, then for any $\mathcal{G}$, $n \geq 0$, $\epsilon > 0$,*

$$\Pr\left(\sup_{G \in \mathcal{G}} |\widehat{Q}(G) - Q(G)| \geq \epsilon\right) \leq 8 S_{\mathcal{G}}(n) e^{-n\epsilon^2/32}$$

where $\widehat{Q}(G) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \in G\}}$.

Proof. Define P_{XY} on $X \times \{0,1\}$ s.t. $P_{XY}(Y = 0) = 1$,$P_{X|Y=0} = Q$, and $P_{X|Y=1}$ is arbitrary. Then

$$\begin{aligned} R(h) &= P_{XY}(h(X) \neq Y) \\ &= P_{XY}(Y = 1) P_{X|Y=1}(h(X) = 0) + P_{XY}(Y = 0) P_{X|Y=0}(h(X) = 1) \\ &= Q(G_h) \end{aligned}$$

Similarly, $\widehat{R}_n(h) = \widehat{Q}(G_h)$, and so

$$\sup_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| = \sup_{G \in \mathcal{G}} |\widehat{Q}(G) - Q(G)|$$

where $\mathcal{H}$ is defined in terms of $\mathcal{G}$ via $G \mapsto h_G = \mathbf{1}_{\{x \in G\}}$. Finally, not that $S_{\mathcal{G}}(n) = S_{\mathcal{H}}(n)$. □

As an application, consider $X \in \mathbb{R}$ and $\mathcal{G} = \{(-\infty, t] : t \in \mathbb{R}\}$. Then $S_{\mathcal{G}}(n) = n + 1$ (see Fig. 4).

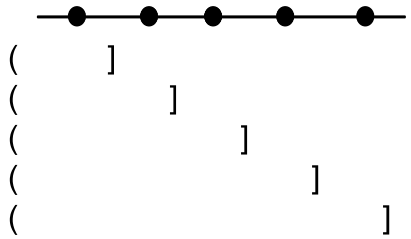

Figure 5: Different ways to intersect sets with points.

Let $X \sim Q$. Denote $G_t = (-\infty, t]$. Then

$$\begin{aligned} Q(G_t) &= \Pr(X \leq t) =: F(t) \text{ (CDF)} \\ \hat{Q}(G_t) &= \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \leq t\}} =: \widehat{F}(t) \text{ (empirical CDF)} \end{aligned}$$

Corollary 7. *For all Q, $n \geq 1$, $\epsilon > 0$,*

$$\Pr(\underbrace{\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)|}_{||\widehat{F}_n - F||_\infty} \geq \epsilon) \leq 8(n+1) e^{-n\epsilon^2/32}.$$

This is known as the Dvoretzky-Kiefer-Wolfowitz inequality. Tighter versions exist [3, 2].

5 Monotone Layers and Convex Sets

Earlier, we obtained convergence guarantees for an empirical risk minimizing classifier when the VC dimension of the classifier set $\mathcal{H}$ was finite. The purpose of this section is to obtain such guarantees also in cases of infinite VC dimension. This requires assumptions about the underlying distribution. In other words the result will not be distribution free, as was possible before. In particular, we will assume that the samples are drawn from a distribution that has a bounded density with compact support. This material is based on a similar result in Chapter 13 of [1], where a weaker assumption on the distribuiton of X is made.

In the following, we consider the feature space $\mathcal{X} = \mathbb{R}^2$ and the following two families of classifiers:

$$\begin{aligned} \mathcal{C} &= \left\{ \mathbf{1}_{\{x \in C\}} \mid C \text{ is convex} \right\}, \\ \mathcal{L} &= \left\{ \mathbf{1}_{\{x \in L\}} \mid L = \{(x_1, x_2) \in \mathbb{R} : x_2 \leq \psi(x_1)\} \text{ for non-increasing } \psi : \mathbb{R} \to \mathbb{R} \right\}. \end{aligned}$$

The classifiers in $\mathcal{L}$ are called *monotone layers.* Both families have infinite VC dimension. The fact that the VC dimension of $\mathcal{C}$ is infinity has been shown earlier. To see that $\mathcal{L}$ also has infinite VC dimension, it suffices to see that any set of points placed decreasingly can be shattered by $\mathcal{L}$, as shown in Figure 6.

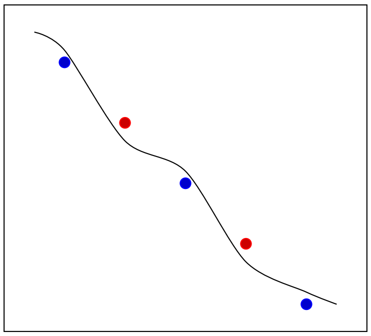

Figure 6: A non-increasing function shatters a set of non-increasingly placed points.

In the proof of the VC theorem, it is shown in an intermediate step that

$$\Pr\left(\sup_{h \in \mathcal{H}} \left| \hat{R}_n(h) - R(h) \right| \geq \varepsilon \right) \leq 8\mathbb{E}\left\{ N_{\mathcal{H}}(X_1, \ldots, X_n) \right\} e^{-n\varepsilon^2/32}.$$

Here, $N_{\mathcal{H}}(X_1, \ldots, X_n)$ is the number of possible labelings of $X_1, \ldots, X_n$ by $\mathcal{H}$. We will now bound this quantity.

Theorem 3. *If X has a density f such that $\|f\|_\infty \leq \infty$ and $\operatorname{supp}(f)$ is bounded, then for $\mathcal{H} = \mathcal{C}$ or $\mathcal{H} = \mathcal{L}$,*

$$\mathbb{E}\left\{ N_{\mathcal{H}}(X_1, \ldots, X_n) \right\} \leq e^{c\sqrt{n}}$$

for a constant c.

Remark. In the theorem statement, $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ is the sup norm. The support of a function is defined as $\operatorname{supp}(f) = \overline{\{x \in \mathcal{X} : f(x) > 0\}}$, where the line denotes the set closure.

Corollary 8. *If $\hat{h}_n$ is an empirical risk minimizer (ERM) over $\mathcal{H} = \mathcal{C}$ or $\mathcal{H} = \mathcal{L}$, then for all $\varepsilon > 0$,*

$$\Pr\left(R(\hat{h}_n) - R^*_{\mathcal{H}} \geq \varepsilon \right) \leq 8e^{c\sqrt{n} - n\varepsilon^2/128}.$$

Equivalently, with probability at least $1 - \delta$,

$$R(\hat{h}_n) - R^*_{\mathcal{H}} \leq \sqrt{\frac{128\left[c\sqrt{n} + \log(8/\delta)\right]}{n}} = \mathcal{O}(n^{-1/4}).$$

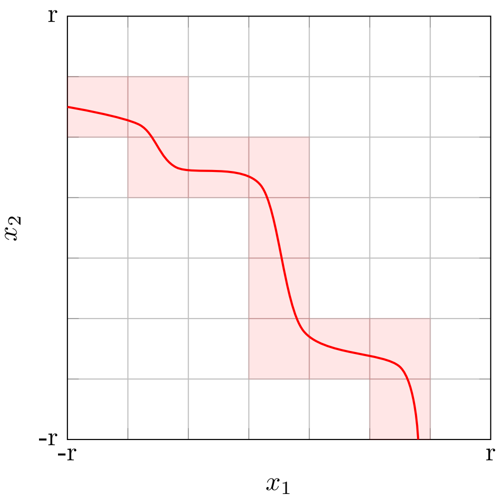

Figure 7: The domain $[-r, r]^2$ is partitioned into m^2 cells, in this case $m = 7$. The red line denotes the points where $x_2 = \psi(x_1)$ and the cells in $c(\psi)$ are marked in red.

Note that the rate of decay is $\mathcal{O}(n^{-1/4})$, whereas when the VC dimension is finite we have a rate of $\mathcal{O}(n^{-1/2})$. We will prove Theorem 3 for the case $\mathcal{H} = \mathcal{L}$; the other case is left as an exercise.

Proof. Fix $r \in \mathbb{N}$ such that $\text{supp}(f) \subseteq [-r, r]^2$. Then divide $[-r, r]^2$ into m^2 squares of side length $2r/m$, as shown in Figure 7.

Let $C_1, \dots, C_{m^2}$ denote the cells in the partition, and let

$$N_j = \#\{i : X_i \in C_j\}$$

be the random variable that counts the number of points in cell j. The random vector $(N_1, \dots, N_{m^2})$ is then multinomially distributed

$$(N_1, \dots, N_{m^2}) \sim \text{multinomial}(n, p_1, \dots, p_{m^2}),$$

where the cell probabilities p_j are

$$p_j = P_X(C_j) = \int\limits_{C_j} f(x)\mathrm{d}x.$$

For a non-increasing $\psi : \mathbb{R} \to \mathbb{R}$, define

$$L(\psi) = \{(x_1, x_2) : x_2 \leq \psi(x_1)\}$$

and

$$c(\psi) = \left\{C_j \mid C_j \text{ intersects both } L(\psi) \text{ and } L(\psi)^C\right\}.$$

In other words, $c(\psi)$ consists of all the cells that contain some part of the graph of ψ, as shown in Figure 7.

Using these definitions, we can bound the number of possible labelings as

$$N_{\mathcal{L}}(X_1, \dots, X_n) \leq \sum_{\text{all possible } c(\psi)} \left(\prod_{C_i \in c(\psi)} 2^{N_i} \right).$$

The sum accounts for all possible non-increasing functions ψ, and the product is an upper bound on the number of labellings for a given $c(\psi)$ obtained by counting the labellings cell-wise. It is clear that cells outside $c(\psi)$ will be uniquely labeled by ψ, therefore only cells in $c(\psi)$ contribute to $N_{\mathcal{L}}$, with obviously at most 2^{N_i} ways to assign labels to the N_i points in C_i.

We now bound the number of terms in the sum and the product separately.

1. It is possible to encode $c(\psi)$ with $2m$ bits. One such encoding can be done using a bit vector
$$(r_1, \ldots, r_{m-1}, c_1, \ldots, c_{m-1}, b_0, b_1) \in \{0,1\}^{2m}.$$
Here the bits are defined as follows:
    - $r_i = 1$ iff the path turns right at row i, $i = 1, \ldots, m-1$.
    - $c_i = 1$ iff the path turns down at column i, $i = 1, \ldots, m-1$.
    - $b_0 = 1$ iff the first turn is right (as opposed to down).
    - $b_1 = 1$ iff the last turn is right (as opposed to down).

   This suffices because paths must alternate down and right turn. Consequently, there are at most 2^{2m} possible $c(\psi)$.

2. We bound the expected value of $\prod_{C_i \in c(\psi)} 2^{N_i}$ using the moment-generating function (MGF) of a multinomial:
$$\mathbb{E}\left\{\prod_{C_i \in c(\psi)} 2^{N_i}\right\} = \mathbb{E}\left\{2^{\sum_{C_i \in c(\psi)} N_i}\right\} = \mathbb{E}\left\{\exp\left(\ln(2) \sum_{C_i \in c(\psi)} N_i\right)\right\}$$
$$= \left(1 + \sum_{C_i \in c(\psi)} p_i\right)^n \leq \exp\left\{n \sum_{C_i \in c(\psi)} p_i\right\}.$$
The last equality follows from the formula for the multinomial MGF, while the last bound follows from
$$\left(1 + \frac{x}{n}\right)^n = \sum_{i=0}^n \binom{n}{i} \left(\frac{x}{n}\right)^i \leq \sum_{i=0}^n \frac{n!}{i!(n-i)!} \left(\frac{x}{n}\right)^i \leq \sum_{i=0}^n \frac{x!}{i!} \leq e^x.$$
Now, since the volume of each cell is $(2r/m)^2$ and there are at most $2m$ cells in a path,
$$\sum_{C_i \in c(\psi)} p_i = \sum_{C_i \in c(\psi)} \int_{C_i} f(x)\, \mathrm{d}x \leq 2m \times \|f\|_\infty \left(\frac{2r}{m}\right)^2 = \frac{8r^2}{m} \|f\|_\infty.$$

By combining 1 and 2, it follows that
$$\mathbb{E}\left\{N_{\mathcal{L}}(X_1, \ldots, X_n)\right\} \leq 2^{2m} e^{8nr^2 \|f\|_\infty / m},$$
a bound which holds for all choices of m. By tuning this parameter as $m \sim \sqrt{n}$, the final bound becomes
$$\mathbb{E}\left\{N_{\mathcal{L}}(X_1, \ldots, X_n)\right\} \leq e^{c\sqrt{n}}$$
for a constant c depending only on r and $\|f\|_\infty$. □

Exercises

1. Determine the sample complexity $N(\epsilon, \delta)$ for ERM over a class $\mathcal{H}$ with VC dimension $V_{\mathcal{H}} < \infty$.

2. Show that the VC Theorem for sets implies the VC Theorem for classifiers. *Hint*: Consider sets of the form $G' = G \times \{0\} \cup G^c \times \{1\} \subset \mathcal{X} \times \mathcal{Y}$, where G^c denotes the complement.

3. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ denote two classes of sets.

   (a) For $\mathcal{G}_1 \cap \mathcal{G}_2 := \{G_1 \cap G_2 \,|\, G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\}$, show $S_{\mathcal{G}_1 \cap \mathcal{G}_2}(n) \leq S_{\mathcal{G}_1}(n) S_{\mathcal{G}_2}(n)$.

   (b) For $\mathcal{G}_1 \cup \mathcal{G}_2 := \{G_1 \cup G_2 \,|\, G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\}$, show $S_{\mathcal{G}_1 \cup \mathcal{G}_2}(n) \leq S_{\mathcal{G}_1}(n) S_{\mathcal{G}_2}(n)$.

4. Show that the following classes have finite VC dimension by exhibiting an explicit upper bound on the VC dimension.

   (a) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathbf{1}_{\{f(x) \geq 0\}} \,|\, f$ is an inhomogeneous quadratic polynomial$\}$.

   (b) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathbf{1}_{\{x \in C\}} \,|\, C$ is a sphere (including boundary and interior)$\}$.

   (c) $\mathcal{X} = \mathbb{R}^2$, $\mathcal{H} = \{\mathbf{1}_{\{x \in P_k\}} \,|\, P_k$ is a convex polygon containing at most k sides$\}$.

   (d) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathbf{1}_{\{x \in R_k\}} \,|\, R_k$ is a union of at most k rectangles$\}$.

5. Prove Theorem 3 for $\mathcal{H} = \mathcal{C}$.

References

[1] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[2] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, 2001.

[3] P. Massart, “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality,” *Annals of Probability*, vol. 18, pp. 1269-1283, 1990.