1 Introduction

So far we have seen some techniques for proving generalization for countably finite hypothesis classes (e.g., union bound), as well as for infinite hypothesis classes (e.g., primal-dual witness, Rademacher complexity).

Example 1. Let’s start by providing a very simple example: classification of one dimensional data. Our task is to give a binary label \{0, 1\} to an input \(z \in \mathbb{R}\). In this example, the hypothesis class is the set of threshold functions:

\[
F = \{ f : \mathbb{R} \to \{0, 1\} \mid f(z) = 1[z > \theta], \theta \in \mathbb{R} \}
\cup
\{ f : \mathbb{R} \to \{0, 1\} \mid f(z) = 1[z < \theta], \theta \in \mathbb{R} \}
\]

Could we use union bounds in this case? The cardinality of \(F\) is equal to the “number”\(^1\) of scalars in \(\mathbb{R}\). Instead of counting the number of functions in \(F\), we will count the possible ways in which \(n\) training samples could be labeled with functions in \(F\).

2 Growth function

We will assume an arbitrary domain \(Z\) and a dataset \(S = \{z_1, \ldots, z_n\}\) containing \(n\) samples where \(z_i \in Z\) for all \(i\). In general, we will assume a hypothesis class \(\mathcal{F} \subseteq \{ f \mid f : Z \to \{0, 1\}\}\). We will use the following shorthand notation:

\[
\mathcal{F}(S) = \{(f(z_1), \ldots, f(z_n)) \in \{0, 1\}^n \mid f \in \mathcal{F}\}
\]

That is, \(\mathcal{F}(S)\) contains all the \(\{0, 1\}^n\) vectors that can produced by applying all functions in \(\mathcal{F}\) to the dataset \(S\).

A natural measure of complexity is the following.

**Definition 7.1.** The growth function (or shatter coefficient) of the hypothesis class \(\mathcal{F} \subseteq \{ f \mid f : Z \to \{0, 1\}\}\) for \(n\) samples is:

\[
G(\mathcal{F}, n) = \max_{S \subseteq Z^n} |\mathcal{F}(S)|
\]

\(^1\)\(\mathbb{R}\) is not a countable set, by Cantor’s diagonalisation argument.
Note that the growth function does not depend on the specific training set $S$, but it is a measure of the worst case among all possible training sets. Clearly $G(F, n) \leq 2^n$, but often it is much smaller.

**Example 1 (continues).** Assume we sort all samples in $S$ in increasing order, and recall that we have threshold functions, thus after sorting it is not possible label three consecutive samples as $0, 1, 0$ or $1, 0, 1$. In other words, all samples to the left should be 0 and all samples to the right should be 1 (or alternatively, all samples to the left should be 1 and all samples to the right should be 0). Let see this more graphically. Each column is one of the $n$ samples, and each row is a possible $\{0, 1\}^n$ vector in $F(S)$.

\[
(1, 0, 0, 0, \ldots, 0) \\
(1, 1, 0, 0, \ldots, 0) \\
(1, 1, 1, 0, \ldots, 0) \\
\vdots \\
(1, 1, 1, 1, \ldots, 1) \\
(0, 1, 1, 1, \ldots, 1) \\
(0, 0, 1, 1, \ldots, 1) \\
\vdots \\
(0, 0, 0, 0, \ldots, 0)
\]

Clearly, $G(F, n) = 2n$ for Example 1.

### 3 Vapnik-Chervonenkis (VC) dimension

**Definition 7.2.** The VC dimension of the hypothesis class $F \subseteq \{ f : Z \rightarrow \{0, 1\} \}$ is:

$$VC(F) = \max_{n \in \mathbb{N}} \{ n \mid G(F, n) = 2^n \}$$

As the growth function, the VC dimension does not depend on the specific training set $S$. Also, by definition if $VC(F) = d$ then for all $n > d$ we have $G(F, n) < 2^n$. (The inequality is strict.) In the following section, we show a less obvious result.

**Example 1 (continues).** As before, assume we sort all samples in $S$ in increasing order, and recall that we have threshold functions. Let's list all possible
\{0,1\}^n \text{ vectors and strikeout the ones that are not in the set } F(S).

<table>
<thead>
<tr>
<th></th>
<th>n = 1</th>
<th>n = 2</th>
<th>n = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0)</td>
<td>(0, 0)</td>
<td>(0, 0, 0)</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(0, 1)</td>
<td>(0, 0, 1)</td>
</tr>
<tr>
<td></td>
<td>(1, 0)</td>
<td>(1, 1)</td>
<td>(0, 1, 1)</td>
</tr>
<tr>
<td></td>
<td>(1, 1)</td>
<td>(0, 1, 0)</td>
<td>(1, 0, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1, 1, 0)</td>
<td>(1, 1, 0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1, 1, 1)</td>
</tr>
</tbody>
</table>

\begin{align*}
G(F, 1) &= 2 \\
G(F, 2) &= 4 \\
G(F, 3) &= 6 
\end{align*}

Clearly, \(VC(F) = 2\) for Example 1. In fact, we previously found that \(G(F, n) = 2n\), by Definition 7.2 we have:

\[
VC(F) = \max_{n \in \mathbb{N}} \{n \mid G(F, n) = 2^n\} \\
= \max_{n \in \mathbb{N}} \{n \mid 2n = 2^n\} \\
= 2
\]

4 Sauer-Shelah lemma

**Lemma 7.1.** The growth function and the VC dimension of a hypothesis class \(F \subseteq \{f \mid f : \mathbb{Z} \to \{0, 1\}\}\) fulfill:

\[
G(F, n) \leq \sum_{i=0}^{VC(F)} \binom{n}{i} \leq (n + 1)^{VC(F)}
\]

*Proof.* The right-hand side is just a consequence of the binomial theorem, thus we will concentrate on the left-hand side. We will use proof by induction. First, define for clarity:

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

Since the binomial coefficient fulfills \(\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}\), it is clear that:

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

We can restate the theorem as follows, \(VC(F) \leq d\) then:

\[
G(F, n) \leq H(n, d)
\]
**Base case.** We show that eq.(2) holds for $n = 1$ and all $d \geq 1$. Since we have only one sample, we have only two possible $\{0, 1\}^1$ vectors (0) and (1), and therefore:

$$G(\mathcal{F}, 1) = |\{(0), (1)\}|$$

$$= 2$$

On the other hand:

$$H(1, d) = \sum_{i=0}^{d} \binom{1}{i}$$

$$= \binom{1}{0} + \binom{1}{1}$$

$$= 2$$

Thus, $G(\mathcal{F}, 1) = H(1, d) = 2$ and eq.(2) holds for $n = 1$ and all $d \geq 1$.

**Inductive step.** Assume that eq.(2) holds for $n - 1$ and all $d \geq 1$, and show that it holds for $n$ and $d$. Fix a dataset $S$ and define:

$$S = \{z_1, z_2, \ldots, z_n\}$$

$$S_2 = \{z_2, \ldots, z_n\}$$

Furthermore, define:

$$\mathcal{F}_2 = \mathcal{F}(S_2)$$

$$\mathcal{F}'_2 = \{\langle f(z_2), \ldots, f(z_n) \rangle \mid f \in \mathcal{F} \text{ such that } (\exists f' \in \mathcal{F}) f'(z_1) = 1 - f(z_1) \text{ and } f'(z_i) = f(z_i) \text{ for } i = 2 \ldots n\}$$

Let see this more graphically. For $\mathcal{F}(S)$, each column is one of the $n$ samples, and each row is a possible $\{0, 1\}^n$ vector. For $\mathcal{F}_2$ and $\mathcal{F}'_2$, each column is one of the $n - 1$ samples, and each row is a possible $\{0, 1\}^{n-1}$ vector. Here $b_i \in \{0, 1\}$.

<table>
<thead>
<tr>
<th>There are three cases</th>
<th>$\mathcal{F}(S)$</th>
<th>$\mathcal{F}_2$</th>
<th>$\mathcal{F}'_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Two vectors in $\mathcal{F}(S)$ match in entries 2 to $n$, but not in entry 1</td>
<td>$(0, b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
</tr>
<tr>
<td>2. A vector in $\mathcal{F}(S)$ is unique in entries 2 to $n$, entry 1 is 0</td>
<td>$(0, b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
</tr>
<tr>
<td>3. A vector in $\mathcal{F}(S)$ is unique in entries 2 to $n$, entry 1 is 1</td>
<td>$(1, b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
<td>$(b_2, \ldots, b_n)$</td>
</tr>
</tbody>
</table>

Let $c_1$, $c_2$ and $c_3$ the number of times case 1, 2 and 3 occur in $S$, respectively.
The next table shows the number of binary vectors in $\mathcal{F}(S)$, $\mathcal{F}_2$ and $\mathcal{F}_2'$.

| Case | |\(\mathcal{F}(S)\)| |\(\mathcal{F}_2\)| |\(\mathcal{F}_2'\)| |
| --- | --- | --- | --- | --- |
| 1. Two vectors in $\mathcal{F}(S)$ match in \(c_1\) | \(2c_1\) | \(c_1\) | \(c_1\) |
| 2. A vector in $\mathcal{F}(S)$ is unique in \(c_2\) | \(c_2\) | \(c_2\) | 0 |
| 3. A vector in $\mathcal{F}(S)$ is unique in \(c_3\) | \(c_3\) | \(c_3\) | 0 |
| Total number of binary vectors | \(2c_1+c_2+c_3\) | \(c_1+c_2+c_3\) | \(c_1\) |

From the above, it is clear that:

\[
|\mathcal{F}(S)| = |\mathcal{F}_2| + |\mathcal{F}_2'| \quad (3)
\]

Recall that Definition 7.2 (VC dimension) depends on powers of 2. Thus from the above, it is clear that if $\text{VC}(\mathcal{F}(S)) \leq d$ then:

\[
\text{VC}(\mathcal{F}_2) \leq d
\]
\[
\text{VC}(\mathcal{F}_2') \leq d - 1
\]

Recall that the number of samples in $S$ is $n$, while the number of samples in $S_2$ is $n - 1$. From eq.(3), the above and eq.(1), we have:

\[
|\mathcal{F}(S)| = |\mathcal{F}_2| + |\mathcal{F}_2'|
\leq H(n - 1, d) + H(n - 1, d - 1)
= H(n, d)
\]

Since the choice of $S$ was arbitrary, the above holds for any dataset $S$, and thus:

\[
G(\mathcal{F}, n) = \max_{S \in \mathcal{Z}^n} |\mathcal{F}(S)|
\leq H(n, d)
\]

Therefore, eq.(2) holds and we prove our claim.

\[\square\]

## 5 Massart lemma and Rademacher complexity

**Lemma 7.2.** Let $A$ be a countably finite subset of $\mathbb{R}^n$. Let $\sigma = \{\sigma_1 \ldots \sigma_n\}$ be $n$ independent Rademacher random variables. We have:

\[
\mathbb{E}_\sigma \left[ \sup_{a \in A} \left( \sum_{i=1}^{n} \sigma_i a_i \right) \right] \leq \sqrt{2 \log |A|} \sup_{a \in A} \|a\|_2
\]
Proof. For any $t > 0$ we have:

$$
\exp \left( t \mathbb{E}_{\sigma} \left[ \sup_{a \in A} \left( \sum_{i=1}^{n} \sigma_i a_i \right) \right] \right) \leq \mathbb{E}_{\sigma} \left[ \exp \left( t \sup_{a \in A} \left( \sum_{i=1}^{n} \sigma_i a_i \right) \right) \right] \quad (4.\text{a}) 
$$

$$
= \mathbb{E}_{\sigma} \left[ \exp \left( t \sum_{i=1}^{n} \sigma_i a_i \right) \right] 
\leq \mathbb{E}_{\sigma} \left[ \sum_{a \in A} \left( \exp \left( t \sum_{i=1}^{n} \sigma_i a_i \right) \right) \right] 
= \sum_{a \in A} \mathbb{E}_{\sigma} \left[ \exp \left( t \sum_{i=1}^{n} \sigma_i a_i \right) \right] 
= \sum_{a \in A} \prod_{i=1}^{n} \mathbb{E}_{\sigma_i} \left[ \exp \left( t \sigma_i a_i \right) \right] 
= \sum_{a \in A} \prod_{i=1}^{n} \exp \left( \frac{1}{2} t^2 \sigma_i^2 \right) \quad (4.\text{b}) 
= \sum_{a \in A} \exp \left( \frac{1}{2} t^2 \sup_{a \in A} \|a\|_2^2 \right) 
\leq |A| \exp \left( \frac{1}{2} t^2 \sup_{a \in A} \|a\|_2^2 \right)
$$

where the step in eq.(4.a) follows from Jensen’s inequality. The step in eq.(4.b) follows since for all $z \in \mathbb{R}$ we have that $(e^z + e^{-z})/2 \leq e^{z^2/2}$.

By taking logarithms and dividing by $t$ on both sides of the above, we have:

$$
\mathbb{E}_{\sigma} \left[ \sup_{a \in A} \left( \sum_{i=1}^{n} \sigma_i a_i \right) \right] \leq \frac{\log |A|}{t} + \frac{1}{2} t \sup_{a \in A} \|a\|_2^2
$$

In order to minimize the function $f(t) = \frac{\log |A|}{t} + \frac{1}{2} t \sup_{a \in A} \|a\|_2^2$, we make the derivative equal to zero and solve for $t$. That is:

$$
0 = \frac{\partial f(t)}{\partial t} = \frac{-\log |A|}{t^2} + \frac{1}{2} \sup_{a \in A} \|a\|_2^2
$$

Thus, $t = \frac{\sqrt{2 \log |A| \sup_{a \in A} \|a\|_2^2}}{\sup_{a \in A} \|a\|_2}$. Plugging this back in the above, we prove our claim. $\square$
Lemma 7.3. Let $\mathcal{F} \subseteq \{ f \mid f : \mathbb{Z} \to \{0, 1\} \}$ be a hypothesis class. The empirical Rademacher complexity (Definition 5.2) of the hypothesis class $\mathcal{F}$ with respect to $n$ samples is bounded as follows:

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log G(\mathcal{F}, n)}{n}}$$

Proof.

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \right) \right] = \frac{1}{n} \mathbb{E}_\sigma \left[ \sup_{a \in \mathcal{F}(S)} \left( \sum_{i=1}^{n} \sigma_i a_i \right) \right] \leq \frac{1}{n} \sqrt{2 \log |\mathcal{F}(S)|} \sup_{a \in \mathcal{F}(S)} \|a\|_2 \quad (5.a)$$

$$\leq \frac{1}{n} \sqrt{2 \log G(\mathcal{F}, n)} \sqrt{n} \quad (5.b)$$

$$= \sqrt{\frac{2 \log G(\mathcal{F}, n)}{n}} \quad (5.c)$$

where the step in eq.(5.a) follow from Definition 5.2 respectively. The step in eq.(5.b) follows from Massart lemma (Lemma 7.2). The step in eq.(5.c) follows since $G(\mathcal{F}, n)$ and since for all $a \in \{0, 1\}^n$ we have that $\|a\|_2 \leq \sqrt{n}$. \qed