1 Theorem 3.2 (continued from lecture #4)

In general, we are trying to show that, with probability \( \geq 1 - \delta \) for all \( h \) in our hypothesis space, that \( h \) being consistent implies \( \text{err}_D(h) \leq \epsilon \). To do this we are bounding the probability that there exists an \( h \) such that \( h \) is consistent yet has error \( \geq \epsilon \).

1.1 Review of Previous Results

We were in the middle of proving that, with probability \( \geq 1 - \delta \), \( \forall h \in H \):

\[
\quad \text{consistent} \Rightarrow \text{err}_D(h) \leq O\left(\frac{\ln \Pi_H(2m) + \ln \frac{1}{\delta}}{m}\right)
\]

The following has been established (or asserted and deferred):

\[
\Pr[B] \leq 2 \Pr[B'] \quad \text{(2)}
\]

\[
\Pr[e(h)|S,S'] \leq 2^{-m\epsilon_2} \quad \text{(3)}
\]

Where:

\[
S \equiv \text{our training sample of } m \text{ random points according to } D
\]

\[
S' \equiv \text{our other sample of } m \text{ random points according to } D
\]

\[
M(h) \equiv \text{the number of mistakes } h \text{ makes on } S'
\]

\[
e(h) \equiv \text{h consistent with } S \land M(h) \geq \frac{me}{2}
\]

\[
B \equiv \exists h \in \mathcal{H} : h \text{ consistent with } S \land \text{err}_D(h) > \epsilon
\]

\[
B' \equiv \exists h \in \mathcal{H} : e(h)
\]

1.2 Working with Fixed \( S, S' \)

Let \( \mathcal{H}' \equiv \{\text{one representative from } \mathcal{H} \text{ for every dichotomy of } S; S'\} \). Clearly, we have another interpretation of \( B' \):

\[
B' \equiv \exists h \in \mathcal{H}' : e(h)
\]

If we call the elements of \( \mathcal{H}' h_1, h_2, \ldots, h_N \), we can then use the union bound:

\[
\Pr[B'|S,S'] = \Pr[\exists h \in \mathcal{H}' : e(h)|S,S']
\]

\[
= \Pr[e(h_1) \lor e(h_2) \lor \ldots \lor e(h_N)|S,S']
\]

\[
\leq \sum_{i=1}^{N} \Pr[e(h_i)|S,S']
\]

\[
\leq |\mathcal{H}'| \cdot 2^{-m\epsilon_2}
\]

\[
= |\Pi_{\mathcal{H}}(S;S')| \cdot 2^{-m\epsilon_2}
\]
1.3 Unfixing Variables in General

We now take a break from the proof to explore a method for eliminating our dependance on a fixed $S$ and $S'$. Let $A$ be an arbitrary event, and $X$ a random variable (it is irrelevant whether or not $X$ and $A$ are independent). Well, by the definitions of probability (see the notes from lecture #2),

$$
Pr[A] = \sum_x Pr[A \land X = x]
$$

(16)

$$
= \sum_x Pr[X = x] \cdot Pr[A|X = x]
$$

(17)

$$
= E_X [Pr[A|X]]
$$

(18)

1.4 Unfixing $S$ and $S'$ and Completing the Proof

Now we can use this result to bound $Pr[B']$ with our bound for $Pr[B'|S,S']$:

$$
Pr[B'] = E_{S,S'} [Pr[B'|S,S']]
$$

(19)

$$
\leq E_{S,S'} [\Pi_H(S; S') \cdot 2^{-\frac{m\epsilon}{2}}]
$$

(20)

$$
\leq E_{S,S'}[\Pi_H(2m) \cdot 2^{-\frac{m\epsilon}{2}}]
$$

(21)

$$
= \Pi_H(2m) \cdot 2^{-\frac{m\epsilon}{2}}
$$

(22)

Using our other previous result:

$$
Pr[B] \leq 2Pr[B']
$$

(23)

$$
\leq 2\Pi_H(2m) \cdot 2^{-\frac{m\epsilon}{2}}
$$

(24)

Finally, setting this bound $\leq \delta$, we find that, with probability $\geq 1 - \delta, \forall h \in H$,

$$
err_D(h) \leq \epsilon \leq \frac{2 \cdot (\log \Pi_H(2m) + \frac{1}{\delta} + 1)}{m}
$$

(25)

2 The VC Dimension

The result we just derived is, of course, completely useless if we can’t bound $\Pi_H(2m)$ to some sub-exponential order, with respect to $m$. Sauer’s lemma will do just that, but first we need to explore a new concept: the Vapnik-Chervonenkis Dimension.

2.1 Definitions

$S$ is said to be shattered by $H$ if every dichotomy of $S$ has a representative in $H$ (i.e. $|\Pi_H(S)| = 2^{|S|}$).

The VC dimension of $H$ is defined to be the size of the largest $S$ which is shattered by $H$ (i.e. $vc(H) = \max \{|S| : S \text{ is shattered by } H\}$)
2.2 Example: Intervals in $\mathbb{R}$

For example, let $\mathcal{H} = \{\text{intervals in } \mathbb{R}\}$. When $S$ is composed of 1 or 2 samples, $S$ is quite obviously shattered. If follows that $VCdim(\mathcal{H}) \geq 2$.

![Figure 1: Representatives of $\mathcal{H}$ which shatter $S$ when $S$ is a set of 1 or 2 points.](image)

However, when $S$ is composed of 3 sample points, it is not shattered (if our sample points are $x_1$, $x_2$, and $x_3$ with $x_1 < x_2 < x_3$, there is no hypothesis which can label just $x_1$ and $x_3$ positive without also labelling $x_2$ positive).

![Figure 2: When $S$ is a set of 3 points, we cannot find a hypothesis which marks the two outer points positive without also marking the inner point so.](image)

To show that $VCdim(\mathcal{H}) < 3$, it is not sufficient to show that a single set of size 3 is not shattered. We need to show that no set of size 3 is shattered. However, in this case, it is evident that our argument applies to all sets of size 3. Thus, $VCdim(\mathcal{H}) = 2$.

Note that if no set of size $d$ is shattered, then no larger set can be shattered either.
2.3 Example: Rectangles $\mathbb{R}^2$

Let $\mathcal{H} = \{\text{rectangles in } \mathbb{R}^2\}$. We will use a proof by picture to show that there is an $S$ such that $|S| = 4$ and $S$ is shattered by $\mathcal{H}$:

![Diagram of rectangles shattering S](image)

Figure 3: Representatives of $\mathcal{H}$ which shatter $S$ when $S$ is a set of 4 points.

Thus $VCdim(\mathcal{H}) \geq 4$, now we need to show that $VCdim(\mathcal{H}) < 5$.

Suppose $|S| \geq 5$. If you take the leftmost, rightmost, topmost, and bottommost points of $S$, there is at least one other point, and it must logically be inside. As such, no rectangle can label the leftmost, rightmost, topmost, and bottommost points of $S$ positive without also labeling the interior point positive.

2.4 The VC Dimension of Finite Hypothesis Spaces

Since each hypothesis corresponds to precisely one dichotomy of $S$, the number of dichotomies of $S$ is less than or equal to $|\mathcal{H}|$. Furthermore, since a shattered $S$ requires $2^{|S|}$ dichotomies,

$$2^{VCdim(\mathcal{H})} \leq |\mathcal{H}|$$  \hspace{1cm} (26)
So,

$$|VCdim(\mathcal{H})| \leq \log |\mathcal{H}|$$  \hspace{1cm} (27)

### 2.5 Sauer’s Lemma

*Sauer’s Lemma* states that,

$$\Pi_d(m) \leq \Phi_d(m)$$  \hspace{1cm} (28)

Where:

$$d \equiv VCdim(\mathcal{H})$$  \hspace{1cm} (29)

$$\Phi_d(m) \equiv \sum_{i=0}^{d} \binom{m}{i}$$  \hspace{1cm} (30)

### 2.6 The Proof of Sauer’s Lemma

Note that it is a common convention that, \( \binom{n}{k} \equiv 0 \) if \( k < 0 \) or \( k > n \). In our proof, we shall also use the following proposition, which turns out to be true even with the aforementioned convention:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$  \hspace{1cm} (31)

The following proof will be done by induction on \( (m + d) \):

**Base Cases:**

Whenever \( d = 0 \), \( \mathcal{H} \) can’t even shatter an \( S \) of one point. Thus all \( h \in \mathcal{H} \) label all points the same way (whether it be positive or negative). Thus, all the \( h \) are identical and \( |\mathcal{H}| = 1 \). So regardless of \( m \), \( \Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0} = \Phi_0(m) \).

On the other hand, whenever \( m = 0 \), there is only one way to label a set of 0 examples. Thus, regardless of \( \mathcal{H} \), \( \Pi_{\mathcal{H}}(0) = 1 = \binom{0}{0} + \binom{0}{1} + \ldots + \binom{0}{d} = \Phi_d(0) \).

**Induction Hypothesis:**

Assume the lemma to be true for all \( m' \) and \( d' \) in which \( m' + d' < m + d \).

**Induction Step:**

Let us work on \( m \) sample points, \( S = \{x_1, x_2, \ldots, x_m\} \), with a hypothesis space \( \mathcal{H} \) of VC dimension \( d \), \( VCdim(\mathcal{H}) = d \). For convenience, let \( S_{\setminus m} = \{x_1, x_2, \ldots, x_{m-1}\} \).

We define two new (finite) hypotheses spaces, \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \), in the following manner:
\[ \mathcal{H}_0 \equiv \{ \text{one representative from } \mathcal{H} \text{ for each dichotomy over } S \} \quad (32) \]

\[ \mathcal{H}_1 \equiv \{ \text{one representative from } \mathcal{H}_0 \text{ for each dichotomy over } S \setminus m \} \quad (33) \]

\[ \mathcal{H}_2 \equiv \mathcal{H}_0 - \mathcal{H}_1 \quad (34) \]

Take, for example, some \( \mathcal{H} \) which contains the dichotomies given by the \( \mathcal{H}_0 \) column of the table below, where \( m = 4 \). The following table illustrates the procedure (hypotheses are identified by their dichotomies for the sake of readability):

<table>
<thead>
<tr>
<th>( \mathcal{H}_0 )</th>
<th>( \mathcal{H}_1 )</th>
<th>( \mathcal{H}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>01100</td>
<td>( \rightarrow )</td>
<td>0110</td>
</tr>
<tr>
<td>01101</td>
<td>( \rightarrow )</td>
<td>01101</td>
</tr>
<tr>
<td>01110</td>
<td>( \rightarrow )</td>
<td>0111</td>
</tr>
<tr>
<td>10100</td>
<td>( \rightarrow )</td>
<td>1010</td>
</tr>
<tr>
<td>10101</td>
<td>( \rightarrow )</td>
<td>10101</td>
</tr>
<tr>
<td>11001</td>
<td>( \rightarrow )</td>
<td>1100</td>
</tr>
</tbody>
</table>

So for \( \mathcal{H}_1 \) over \( S \setminus m \), \( m_1 = m - 1 \) (because \( S \) is one smaller than \( S \setminus m \)) and \( d_1 = VCdim(\mathcal{H}_1) \leq d \) (because reducing the number of hypotheses certainly will not increase the VC dimension of a space).

Similarly, with \( \mathcal{H}_2 \) over \( S \setminus m \), \( m_2 = m - 1 \) and \( d_2 = VCdim(\mathcal{H}_2) \leq d - 1 \). Let us explain \( d - 1 \): By construction, if \( S' \subseteq S \setminus m \) is shattered by \( \mathcal{H}_2 \), then every dichotomy over \( S' \) must occur both in \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) but with different labelings of \( x_m \). Thus, \( S' \cup \{ x_m \} \), which has size \( |S'| + 1 \), is shattered by \( \mathcal{H} \), and so \( |S'| \) cannot be more than \( d - 1 \).

Using induction, \( \Pi_{\mathcal{H}_1}(S \setminus m) \leq \Phi_d(m - 1) \) and \( \Pi_{\mathcal{H}_2}(S \setminus m) \leq \Phi_{d-1}(m - 1) \).

Now, by the construction of \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \), \( \Pi_{\mathcal{H}}(S) = |\mathcal{H}_1| + |\mathcal{H}_2| = \Pi_{\mathcal{H}_1}(S \setminus m) + \Pi_{\mathcal{H}_2}(S \setminus m) \). So, using our inequalities along with the convention and proposition put forth at the beginning of this subsection,

\[ \Pi_{\mathcal{H}}(S) = \Pi_{\mathcal{H}_1}(S \setminus m) + \Pi_{\mathcal{H}_2}(S \setminus m) \quad (35) \]

\[ \leq \Phi_d(m - 1) + \Phi_{d-1}(m - 1) \quad (36) \]

\[ = \sum_{i=0}^{d} \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i} \end{array} \right) + \sum_{i=0}^{d-1} \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i} \end{array} \right) \quad (37) \]

\[ = \sum_{i=0}^{d} \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i} \end{array} \right) + \sum_{i=0}^{d} \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i - 1} \end{array} \right) \quad (38) \]

\[ = \sum_{i=0}^{d} \left[ \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i} \end{array} \right) + \left( \begin{array}{c} m - 1 \, \frac{m - 1}{i - 1} \end{array} \right) \right] \quad (39) \]

\[ = \sum_{i=0}^{d} \left( \begin{array}{c} m \, \frac{m}{i} \end{array} \right) \quad (40) \]

\[ = \Phi_d(m) \quad (41) \]
2.7 Sauer’s Lemma and Theorem 3.2

We note that:

\[
\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \tag{42}
\]

\[
= \sum_{i=0}^{d} \frac{m!}{i! \cdot (m - i)!} \tag{43}
\]

\[
= \sum_{i=0}^{d} \frac{(m - 0)(m - 1)(m - 2) \ldots (m - i + 1)}{i!} \tag{44}
\]

\[
= O(m^d) \tag{45}
\]

Thus, \(\Pi_d(m) \leq O(m^d)\), and we have just made Theorem 3.2 useful.