1 First thought

We have \( \Pi_H(S) = \{\langle h(X_1), h(X_2), \ldots, h(X_m) \rangle : h \in H \} \), where \( S = \langle X_1, \ldots, X_m \rangle \). And \( \Pi_H(m) = \max_{S:|S|=m} |\Pi_H(S)| \).

We say that \( H \) shatters \( S \) if \( |\Pi_H(S)| = 2^m \). VC-dim(\( H \)) = max\{\|S\| : H \) shatters \( S \}\). If \( |H| < \infty \), then \( d = \text{VC-dim}(H) \leq \log |H| \). In fact, there are only two cases:

- VC-dim = \( \infty \) \( \Rightarrow \) \( \Pi_H(m) = 2^m \), \( \forall m \)
- VC-dim = \( d < \infty \) \( \Rightarrow \) \( \Pi_H(m) = O(m^d) \)

This follows from Sauer’s Lemma, which we now state and prove.

2 Sauer’s Lemma

**Lemma**: \( \forall H \) with \( d = \text{VC-dim}(H) \),

\[
\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m) = O(m^d).
\]

In other words, the sum of the binomial is just the number of different ways of choosing at most \( d \) items from a set of size \( m \).

2.1 The Interval Example

In our examination of intervals, we found that the equation for the number of dichotomies possible was of the form:

\[
\Pi_H(m) = \binom{m}{2} + \binom{m}{1} + \binom{m}{0} = \Phi_2(m).
\]

So Sauer’s Lemma is tight in this example.

2.2 Proof of Sauer’s Lemma

First, a few facts and conventions will be used in the proof:

- \( \binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \)
- \( \binom{m}{k} = 0 \), if \( k < 0 \) or \( k > m \)

We will prove Sauer’s Lemma by induction on \( m + d \).

**Base cases**: 
Our 2 base cases (for our 2 variables) are:
<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$h_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_5$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_6$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Example Datasets for Proof of Sauers Lemma

- $m = 0$: $\Pi_H(m) = 1 = \sum_{i=0}^{d} \binom{0}{i}$. It is the degenerate labeling of the empty set.
- $d = 0$: $\Pi_H(m) = 1 = \binom{m}{0}$. You can not even shatter one point, so only one behavior possible.

**Inductive Step:**
Assuming lemma holds for any $m' + d' < m + d$. Given $S = \langle x_1, x_2, \ldots, x_m \rangle$, we want to show $|\Pi_H(S)| \leq \Phi_d(m)$.

The main step of the proof is the construction of two new hypothesis spaces: $H_1$ and $H_2$ to which we can apply our induction hypothesis. Here, we have $H_1$ and $H_2$ defined on $S' = X' = \{x_1, x_2, \ldots, x_{m-1}\}$, that is, on all the points except $x_m$. $H_1$ is constructed by just ignoring behavior on $x_m$. $H_2$ is constructed by including only dichotomies that "collapsed" in $H_1$.

As shown in the example in Table 1, $h_1$ and $h_2$, $h_4$ and $h_5$ are the same if we ignore $x_5$, so in each of these pairs, only one of goes to $H_1$, and the other one goes to $H_2$.

Notice that if a set is shattered by $H_1$, then it is also shattered by $H$. The reason is that we can generate $H$ by using the same $x_i$’s when we generate $H_1$. Thus we have

$$\text{VC-dim}(H_1) \leq \text{VC-dim}(H) = d$$

If a set $T$ is shattered by $H_2$, then $T \cup \{x_m\}$ is shattered by $H$ since there will be two corresponding hypotheses in $H$ with each element of $H_2$ by adding $x_m = 1$ and $x_m = 0$. Thus, $\text{VC-dim}(H) \geq \text{VC-dim}(H_2) + 1$, which implies

$$\text{VC-dim}(H_2) \leq d - 1.$$  

Now, by induction, we have:

$$|H_1| = |\Pi_{H_1}(S')| \leq \Phi_d(m - 1).$$

$$|H_2| = |\Pi_{H_2}(S')| \leq \Phi_{d-1}(m - 1).$$
Then, we have
\[
|\Pi_{\mathcal{H}}(S)| = |\mathcal{H}_1| + |\mathcal{H}_2|
\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i}
= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1}
= \sum_{i=0}^{d} \binom{m}{i}
= \Phi_d(m).
\]

### 2.3 Upperbound on $\Phi_d(m)$

**Claim:** $\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$ for $m \geq d \geq 1$.

**Proof:**

\[
\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i}
\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i 1^{m-i}
= \left(1 + \frac{d}{m}\right)^m
\leq e^d.
\]

Then we have $\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$.

Using this bound, we will have the following results:

With probability of at least $1 - \delta$, $\forall h \in \mathcal{H}$, if $h$ is consistent with $m$ examples, then
\[
\text{err}(h) \leq \frac{2}{m} \left[d \log \left(\frac{em}{d}\right) + \log \left(\frac{1}{\delta}\right) + 1\right].
\]

If $m = O\left(\frac{1}{\epsilon^2} [\ln(\frac{1}{\delta}) + d \ln(\frac{1}{\delta})]\right)$, we have $\text{err}(h) \leq \epsilon$.

### 3 About the Lower Bound

Now, let’s try to give a lower bound.

#### 3.1 (Bogus) Argument on Lower Bound

Let $D$ be uniform on $z_1, z_2, \cdots, z_d$. We run $A$ with $m = d/2$ examples labeled arbitrarily, say $A$ outputs $h_A$. Now let $c \in \mathcal{C}$ be any concept that is consistent with labels in $S$ such that $c(x) \neq h_A(x)$ for $x \notin S$. Then we have $\text{err}(h_A) \geq 1/2$.

But, this is not a valid argument because we cannot choose target concept $c$ after we choose $h_A$. The PAC model requires that we choose $c$ before we choose $S$. So, in this argument, we are making $c$ a function of $h_A$, which is in turn a function of $S$, which is obviously wrong.
3.2 A Theorem on the Lower Bound

We will instead prove the following:

**Theorem:** \( \forall A, \exists c \in C, \exists D, \) such that if \( A \) gets \( m = d/2 \) examples, where \( d = \text{VC-dim}(C) \), then

\[
\Pr\left[\text{err}(h_A) > \frac{1}{8}\right] \geq \frac{1}{8}
\]

This means that if given only \( d/2 \) examples, then PAC learning is impossible for \( \epsilon \leq 1/8 \) and \( \delta \leq 1/8 \).