Primal Shatter Function $\Pi_S(n)$

For a set system $S = (X, \mathcal{S})$ of VC-dimension $d$, we know that a set $A$ of $\leq d$ points can be shattered into all $2^n$ unique subsets by intersection with sets in $\mathcal{S}$. What is the maximum number of unique subsets that can be obtained for a set $A$ of size $n > d$? The number of possible subsets as a function of $n$ is known as the Primal Shatter Function and is denoted by $\Pi_S(n)$. The function will have value $2^n$ until $n = d$, at which point it will grow at some slower pace.

Example: Consider the case of rectangles. We can shrink any rectangle until all of its sides contain a point inside, and consider this to be a representative canonical rectangle for any rectangle that contains the same points. A rectangle is defined by at most 4 points that define its edges. Therefore we can choose from at most $n$ points to represent the left side of the rectangle, the right side, etc. Thus there are at most $n^4$ unique canonical rectangles.

We claim that for a set system of finite VC-dimension, the function representing the number of subsets that can be obtained is polynomial in $n$. More specifically:

Lemma: For any set system $S = (X, \mathcal{S})$ with VC-dimension $d$, an upper bound on $\Pi_S(n)$ is given by

$$\Pi_S(n) \leq \sum_{i=0}^{d} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$$

Proof: We will prove this by induction on $n$, and then for fixed $n$ by induction on $d$. Each entry of the table $(n,d)$ will depend on the value of $(n-1,d)$ and $(n-1,d-1)$. Thus we need to fill in the upper part of the chart and the left column as base cases.

Base Case for $n$ ($n \leq d$): We know that $\Pi_S(n) = 2^n$ because our set system can shatter $n$ points. Further, we have $\sum_{i=0}^{d} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$, so this base case holds.
Base Case for \( d = 0 \): We are attempting to determine the value of \( \Pi_S(k) \) where \( S \) has a VC-dimension of 0. What does this mean? For one thing, a set system with VC-dimension 0 must have at most one set. If it had two different sets \( C \) and \( D \), then one of those sets must have 1 point \( x \) which is not in the other. Then we could pick that point \( x \) and the sets \( C \) and \( D \) would shatter it, meaning the set system at least has VC-dimension 1. So a set system of VC-dimension 0 can at most only shatter the empty set. We have \( \Pi_S(n) = 1 \) and \( \sum_{i=0}^{\binom{n}{i}} = 1 \), so this base case holds.

Inductive Step: To bound \( \Pi_S(n) \) for the set system \( S = (X, S) \), we can remove one element \( x \) from \( X \) and consider \( \Pi_{S_1}(n - 1) \) for the system

\[
S_1 = (X - \{x\}, S)
\]

Let \( S \) be some element of \( S \) that does not include \( x \) (we will write \( S \cup \{x\} \) if we want to include it). There are two cases to consider:

Case 1 – exactly 1 of \( S \) and \( S \cup \{x\} \) is in \( S \): Here, we will have exactly 1 of \( S \cap A \) or \( S \cup \{x\} \cap A \) as one of our partitions. We can identify either one of these with the set \( S \cap A_i \) in \( S_2 \), where \( A_i \) is the set \( A \) without the point \( x \). This says that \( \Pi_S(n) \) is equal to \( \Pi_{S_1}(n - 1) \). Since \( \Pi_{S_1}(n - 1) \leq \sum_{i=0}^{\binom{n-1}{i}} \) by the inductive hypothesis, and \( \sum_{i=0}^{\binom{n-1}{i}} \leq \sum_{i=0}^{\binom{n}{i}} \), it must be that \( \Pi_S(n) \leq \sum_{i=0}^{\binom{n}{i}} \) as desired.

Case 2 – both \( S \) and \( S \cup \{x\} \) are in \( S \): In this case, \( S \) and \( S \cup \{x\} \) define distinct subsets in the set system \( S \), but both define the same subset in \( S_1 \). This tells us that \( \Pi_{S_1}(n - 1) \) and \( \Pi_S(n) \) differ by the cardinality of the set \( \{X - \{x\} \cap S \mid both S and S \cup \{x\} are in S\} \). We will define the set

\[
S_2 = (X - \{x\}, \{S \mid S \text{ and } S \cup \{x\} \text{ are in } S\})
\]

Then we have the following recurrence relation:

\[
\Pi_S(n) = \Pi_{S_1}(n - 1) + \Pi_{S_2}(n - 1)
\]

We know bounds on the latter two terms by the inductive hypothesis, so we just have to add them together.

Claim: \( S_1 \) has VC-dimension \( \leq d \).

To see this, suppose that the VC-dimension is \( \geq d \). Then there exists some set \( A \), \( |A| > d \) that can be shattered by \( S \), which is a contradiction.
Claim: $S_2$ has VC-dimension $\leq d-1$.

To see this, note that if $A - \{x\}$ is shattered in $S_2$ then $A$ is shattered in $S$. If $S_2$ had VC-dimension $> d-1$, then $S$ would have VC-dimension $> d$, a contradiction.

By the inductive hypothesis, $\Pi_{S_1}(n-1) \leq \sum_{i=0}^{d} \binom{n-1}{i}$ and $\Pi_{S_2}(n-1) \leq \sum_{i=0}^{d-1} \binom{n-1}{i}$.

Thus

$$\Pi_S(n) = \Pi_{S_1}(n-1) + \Pi_{S_2}(n-1) \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i}$$

$$= \left[\binom{n-1}{0} + \binom{n-1}{1} + \cdots + \binom{n-1}{d-1}\right] + \left[\binom{n-1}{0} + \binom{n-1}{1} + \cdots + \binom{n-1}{d-1}\right]$$

$$= \binom{n-1}{0} + \left[\binom{n-1}{0} + \binom{n-1}{1}\right] + \cdots + \left[\binom{n-1}{d-1} + \binom{n-1}{d}\right]$$

$$= \sum_{i=0}^{d} \binom{n-1}{i} + \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n}{i}$$

We use the fact $\binom{n-1}{i} + \binom{n-1}{i-1} = \binom{n}{i}$. Combinatorial proof: to choose a subset $T$ of size $i$ from a set $S$ of size $n$, pick an element $x$. If $x$ is in $T$, then there are $\binom{n-1}{i-1}$ ways to choose the remaining $i-1$ elements. If $x$ is not in $T$, then there are $\binom{n-1}{i}$ ways to choose the $i$ elements from $S - \{x\}$. 