5.2 Primal Shatter Property and ϵ-nets

> *"Mathematics is one of a few fields in which one can do top-level work without a lot of life experience, something that might be key in the arts or humanities. One does not have to have experience raising children through school, dealing with family tragedies, and so forth, to be able to find three numbers whose fourth powers add up to another one."*
>
> Noam Elkies

Recall (Section 4.3) that for general set systems $(X, \mathcal{F})$, there exist ϵ-nets of size $O(\frac{\log|\mathcal{F}|}{\epsilon})$ and that there are examples for which this is the best possible bound. We now present improved $O(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ bounds for the class of set systems with bounded VC-dimension.

Definition 5.1. *A set system $(X, \mathcal{F})$ satisfies* primal shatter property *with constant d if for all $Y \subseteq X$, $|\mathcal{F}|_Y| = O(|Y|^d)$.*

Observe that by Lemma 5.1, a set system with VC-dimension d satisfies primal shatter property with constant d. The main theorem we will prove in this section is:

Theorem 5.2. *Let $(X, \mathcal{F})$ be a set system where $|X| = n$ and $\mathcal{F} = \{S_1, \ldots, S_m\}$ satisfies the primal shatter property with constant d. Then there exists an ϵ-net for $\mathcal{F}$ of size $O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$.*

Intuitive idea. W.l.o.g. assume that each set in $\mathcal{F}$ has size at least ϵn. Imagine that we could pick a set $R_1 \subseteq X$ such that $|R_1| = \frac{n}{2}$ and R_1 is also 'equally representative' of all the other sets in $\mathcal{F}$. Namely, R_1 contains exactly half of the elements of each set of $\mathcal{F}$: for all $S \in \mathcal{F}$, we have $|R_1 \cap S| = \frac{|S|}{2}$. Repeating the above step, imagine we could pick $R_2 \subseteq R_1$, with $\frac{|R_1|}{2} = \frac{n}{4}$ elements, such that $|R_2 \cap S| = \frac{|R_1 \cap S|}{2} = \frac{|S|}{4}$. See Figure 5.1. Iterating, let R_j be the set at the j-th iteration. Note that

$$|R_j| = \frac{n}{2^j} \quad \text{and} \quad |R_j \cap S| = \frac{|S|}{2^j} \ \text{ for all } S \in \mathcal{F}.$$

Setting $j = \log(\epsilon n)$, we get

$$|R_{\log(\epsilon n)}| = \frac{n}{2^{\log(\epsilon n)}} = \frac{1}{\epsilon} \quad \text{and} \quad |R_{\log(\epsilon n)} \cap S| = \frac{|S|}{2^{\log(\epsilon n)}} = \frac{|S|}{\epsilon n} \geq 1 \ \text{ for all } S \in \mathcal{F},$$

since $|S| \geq \epsilon n$. Hence, $R_{\log(\epsilon n)}$ is an ϵ-net with $O(1/\epsilon)$ elements!

The problem with the above idea is that the existence of a $R_1 \subseteq X$ that contains $\frac{|X|}{2}$ points, and also exactly halves *every* set is not always possible. In fact, how well one can pick R_1 will have to depend on the number of sets in $\mathcal{F}$. For example, for the complete set system $\mathcal{F} = 2^X$, for any choice of R_1 of $\frac{n}{2}$ elements, there will be a set of $\frac{n}{2}$ elements in $\mathcal{F}$, namely the set $X \setminus R_1$, that will not share *any* points of R_1.

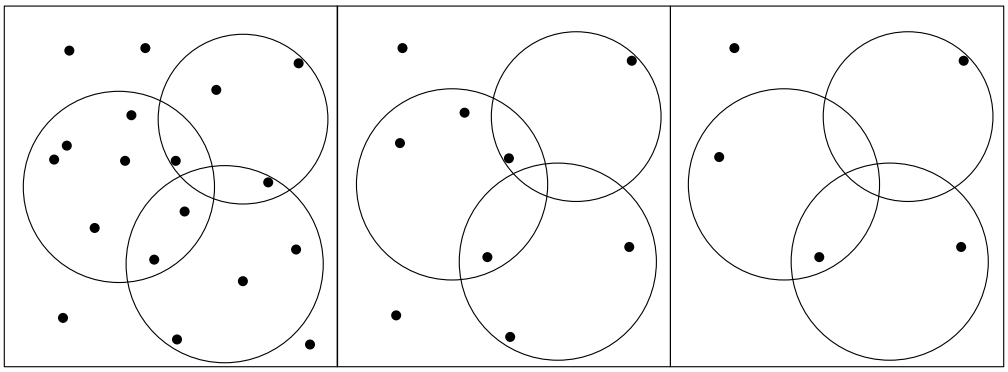

Figure 5.1: At each step, the total number of points halve; ideally, the number of points within each disk is also halved.

Thus one can ask for the best R_1, $|R_1| = \frac{n}{2}$, such that for all $S \in \mathcal{F}$, $|R_1 \cap S|$ is as close to $\frac{|S|}{2}$ as possible, although necessarily as a function of m. Not surprisingly, we will pick R_1 randomly, by adding each element $p \in X$ to R_1 with probability $1/2$. First, note that the expected number of elements picked into R_1 is $n/2$. Not only that, but the expected number of elements picked from any *fixed* $S \in \mathcal{F}$ is also $\frac{|S|}{2}$. Of course that is unlikely to be true for all sets *simultaneously*[†].

The key to proving our main theorem will be the *Halving Claim*, which states the existence of a set R_1, $|R_1| = \frac{n}{2}$, and a function $f(\cdot, \cdot)$ such that:

$$|R_1 \cap S| \geq \frac{|S|}{2} - f(n, m) \quad \text{for all } S \in \mathcal{F},$$

where $f(n, m)$ is called the *error term*. Then, using the Halving Claim, we iteratively construct R_j as before. However, unlike earlier, we cannot continue till $j = \log(\epsilon n)$, as the error terms add up to be too large (i.e., greater than $|S|/2^j$). Instead, we will continue until a smaller value of j, so that the following is true:

$$|R_j| = O\Big(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon}\Big) \text{ and } |R_j \cap S| = \Omega\Big(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\Big) \text{ for all } S \in \mathcal{F}. \tag{5.1}$$

Assume that using the Halving Claim, one can find such a j satisfying the above conditions. So for each $S \in \mathcal{F}$, we have that $S \cap R_j$ contains an ϵ-th proportion of the elements of R_j. Thus an ϵ-net for the set system $(R_j, \mathcal{F}|_{R_j})$ hits every set of $\mathcal{F}$. Finally, observe that by the Primal Shatter Property,

$$|\mathcal{F}|_{R_j}| = O(|R_j|^d) = O\Big(\big(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon}\big)^d\Big).$$

Using the ϵ-net result for general set systems (Theorem 4.15), there exists an ϵ-net for $(R_j, \mathcal{F}|_{R_j})$ of size at most

$$\frac{\log |\mathcal{F}_{R_j}|}{\epsilon} = \frac{\log(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})^d}{\epsilon} = \frac{d}{\epsilon}\big(2 \log \frac{1}{\epsilon} + \log \log \frac{1}{\epsilon}\big).$$

[†]Take the complete set system example again.

Therefore it remains to show the existence of R_j as specified in Equation(5.1).

PROOF OF MAIN THEOREM

We now turn to the formal proof of the main theorem. The precise statement of the Halving claim is the following.

Claim 5.3 (Halving Claim)**.** *Given a set system $(X, \mathcal{F})$, $|X| = n$, $|\mathcal{F}| = m$, there exist constants c, c' and a set $R_1 \subseteq X$, $|R_1| = \frac{n}{2}$, such that*

$$\text{for all } S \in \mathcal{F}, \;\; |R_1 \cap S| > \frac{|S|}{2} - c'\sqrt{n \log m}.$$

If $\mathcal{F}$ satisfies the primal shatter property with constant d, we have $m = O\big(|X|^d\big)$, and so it follows that for some constant $c > 0$, we have

$$|R_1 \cap S| > \frac{|S|}{2} - c\sqrt{n \log |X|^d}.$$

Proof. We prove this claim by showing that a random R_1 suffices with non-zero probability. Fix a number Δ. Independently with probability $1/2$, assign a value '+1' or '-1' to each $p_i \in X$. Let X_i be the value assigned to p_i. By the Chernoff bound:

$$\text{For each } S \in \mathcal{F}, \quad \Pr\left[\sum_{p_i \in S} X_i \geq \Delta\right] \leq e^{-\frac{\Delta^2}{2|S|}} \leq e^{-\frac{\Delta^2}{2n}}.$$

Set $\Delta = \sqrt{2n \log(4m)}$; by the union bound the probability that R_1 fails for at least one set in $\mathcal{F}$ is:

$$\sum_{S \in \mathcal{F}} \Pr\Big[\sum_{p_i \in S} X_i \geq \Delta\Big] \leq \sum_{S \in \mathcal{F}} e^{-\frac{\Delta^2}{2n}} = m \cdot e^{-\frac{\Delta^2}{2n}} \leq \frac{1}{4}.$$

Similarly, the probability that at least one set S has $\sum_{p_i \in S} X_i \leq -\Delta$ is at most $1/4$. Thus with probability at least $1/2$, we have

$$\text{For each } S \in \mathcal{F}, \quad -\Delta \leq \sum_{p_i \in S} X_i \leq \Delta.$$

For each $S \in \mathcal{F}$, let $S^+ \subseteq S$ be the elements with value $+1$, and S^- the elements with value -1. Therefore, with probability at least $1/2$, for each set $S \in \mathcal{F}$, we have

$$-\Delta \leq |S^+| - |S^-| \leq \Delta$$

As $|S^+| + |S^-| = |S|$, this implies that

$$-\Delta \leq 2|S^+| - |S| \leq \Delta \implies \frac{|S|}{2} + \frac{\Delta}{2} \geq |S^+| \geq \frac{|S|}{2} - \frac{\Delta}{2}$$

Coming back to our problem of choosing R_1, pick all the elements assigned '+1' into the set R_1. Now for each set S, we have picked at least $\frac{|S|}{2} - \frac{\Delta}{2}$ elements into R_1. We are done, except a minor technical detail: we need to pick exactly $n/2$ elements[†], which need not necessarily be true for either of '+1' or '-1' valued elements.

Assuming w.l.o.g. that X is also a set in $\mathcal{F}$[‡], and that there are at least $n/2$ elements assigned '+1'; then $\frac{n}{2} \leq |X^+| \leq \frac{n}{2} + \frac{\Delta}{2}$. Pick any $n/2$ elements from X^+ arbitrarily to form our set R_1. Now each set $S \in \mathcal{F}$ still contains at least

$$\Big(\frac{|S|}{2} - \frac{\Delta}{2}\Big) - \frac{\Delta}{2} = \frac{|S|}{2} - \sqrt{2n\log(4m)}$$

elements of R_1. □

Proof of Theorem 5.2. We will iteratively compute R_i. Apply the Halving Claim to $(X, \mathcal{F})$ to get a set $R_1 \subset X$, $|R_1| = \frac{n}{2}$, such that

$$\text{for all } S \in \mathcal{F}: \quad |R_1 \cap S| \geq \frac{|S|}{2} - c\sqrt{n \log n^d}$$

Consider the set system

$$\mathcal{F}_1 = \mathcal{F}|_{R_1} = \{R_1 \cap S \mid S \in \mathcal{F}\}.$$

By the primal shatter property,

$$|\mathcal{F}_1| = O\big(|R_1|^d\big) = O\Big(\big(\frac{n}{2}\big)^d\Big).$$

Now again apply the Halving Claim on $(R_1, \mathcal{F}_1)$ to get a set $R_2 \subset R_1$, $|R_2| = \frac{|R_1|}{2} = \frac{n}{4}$, such that for all $S \in \mathcal{F}$:

$$\begin{aligned}
|R_2 \cap S| &\geq \frac{|R_1 \cap S|}{2} - c\sqrt{|R_1| \log |\mathcal{F}_1|} \\
&\geq \frac{|R_1 \cap S|}{2} - c\sqrt{\frac{n}{2}\log(\frac{n}{2})^d} \\
&\geq \frac{\frac{|S|}{2} - c\sqrt{n \log n^d}}{2} - c\sqrt{\frac{n}{2}\log(\frac{n}{2})^d} \\
&= \frac{|S|}{4} - \frac{c\sqrt{n \log n^d}}{2} - c\sqrt{\frac{n}{2}\log(\frac{n}{2})^d} \\
&= \frac{|S|}{4} - \sum_{j=0}^{1} \frac{c\sqrt{\frac{n}{2^j}\log(\frac{n}{2^j})^d}}{2^{1-j}}.
\end{aligned}$$

In general, assume we have the set system

$$\big(R_{i-1}, \mathcal{F}_{i-1} = \{R_{i-1} \cap S \mid S \in \mathcal{F}\}\big)$$

[†]This condition that $|R_1|$ be exactly equal to $n/2$ is not really necessary ($|R_1| = n/2 \pm \Delta$ would suffice), but it considerably simplifies the subsequent algebraic calculations.

[‡]so that the number of elements with value '+1', and those with value '-1' are within Δ of each other.

where

$$|R_{i-1}| = \frac{n}{2^{i-1}}, \quad |\mathcal{F}_{i-1}| = O(|R_{i-1}|^d) = O\Big((\frac{n}{2^{i-1}})^d\Big).$$

Furthermore, by induction, assume that

$$|R_{i-1} \cap S| \geq \frac{|S|}{2^{i-1}} - \sum_{j=0}^{i-2} \frac{c\sqrt{\frac{n}{2^j}\log(\frac{n}{2^j})^d}}{2^{(i-2)-j}}$$

Applying the Halving Claim to $(R_{i-1}, \mathcal{F}_{i-1})$, we get a set R_i, $|R_i| = \frac{|R_{i-1}|}{2} = \frac{n}{2^i}$, with

$$\begin{aligned}
|R_i \cap S| &\geq \frac{|R_{i-1} \cap S|}{2} - c\sqrt{|R_{i-1}| \log |\mathcal{F}_{i-1}|} \\
&\geq \frac{|R_{i-1} \cap S|}{2} - c\sqrt{\frac{n}{2^{i-1}}\log(\frac{n}{2^{i-1}})^d} \\
&\geq \frac{|S|}{2^i} - \sum_{j=0}^{i-2} \frac{c\sqrt{\frac{n}{2^j}\log(\frac{n}{2^j})^d}}{2^{(i-1)-j}} - c\sqrt{\frac{n}{2^{i-1}}\log(\frac{n}{2^{i-1}})^d} \\
&= \frac{|S|}{2^i} - \sum_{j=0}^{i-1} \frac{c\sqrt{\frac{n}{2^j}\log(\frac{n}{2^j})^d}}{2^{(i-1)-j}}.
\end{aligned}$$

We first bound the second error term in the last equation:

$$\begin{aligned}
&\sum_{j=0}^{i-1} \frac{c\sqrt{\frac{n}{2^j}\log(\frac{n}{2^j})^d}}{2^{(i-1)-j}} \\
= \; &\frac{c\sqrt{dn}}{2^{i-1}} \sum_{j=0}^{i-1} \sqrt{2^j \log(\frac{n}{2^j})} \\
= \; &\frac{c\sqrt{dn}}{2^{i-1}} \cdot O\Big(\sqrt{2^i \log \frac{n}{2^i}}\Big) = \frac{c'\sqrt{n2^i \log(\frac{n}{2^i})}}{2^i}
\end{aligned}$$

We set the number of iterations i so that the error term is negligible, with $i = \log\big(\frac{\epsilon^2 n}{c_1 \log \frac{1}{\epsilon}}\big)$ for a constant c_1 to be set later. Then the error term becomes:

$$\begin{aligned}
&\frac{c'\sqrt{n2^i \log(\frac{n}{2^i})}}{2^i} \\
= \; &c' \cdot \frac{1}{\epsilon} \cdot \sqrt{c_1} \cdot \sqrt{\log \frac{1}{\epsilon} \cdot \log\big(\frac{c_1 \log \frac{1}{\epsilon}}{\epsilon^2}\big)} \\
\leq \; &c' \cdot \frac{1}{\epsilon} \cdot \sqrt{c_1} \cdot 4 \log \frac{c_1}{\epsilon}. \qquad (5.2)
\end{aligned}$$

Now since each $S \in \mathcal{F}$ has size at least ϵn, we have $\frac{|S|}{2^i} \geq \frac{\epsilon n}{2^i} = \frac{c_1 \log 1/\epsilon}{\epsilon}$. For a large enough

constant c_1, this becomes more than twice the error term (5.2). Therefore, we have:

$$\begin{aligned} |R_i \cap S| &\geq \frac{|S|}{2^i} - \Big(c' \cdot \frac{1}{\epsilon} \cdot \sqrt{c_1} \cdot 4 \log \frac{c_1}{\epsilon}\Big) \\ &\geq \frac{c_1 \log \frac{1}{\epsilon}}{\epsilon} - c' \cdot \frac{1}{\epsilon} \cdot \sqrt{c_1} \cdot 4 \log \frac{c_1}{\epsilon} \\ &\geq \frac{c_1 \log \frac{1}{\epsilon}}{2\epsilon}. \end{aligned}$$

So for $i = \log\big(\frac{\epsilon^2 n}{c_1 \log \frac{1}{\epsilon}}\big)$, consider the set system $(R_i, \mathcal{F}|_{R_i})$. We have $|R_i \cap S| = \Omega(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$, where the total number of elements, R_i, are $n/2^i = \frac{c_1 \log \frac{1}{\epsilon}}{\epsilon^2}$. In other words, each set $(R_i \cap S) \in \mathcal{F}|_{R_i}$ contains more than an ϵ-th proportion of the total number of elements, $|R_i|$. By the ϵ-net bound for general set systems, there exists a set $R' \subseteq R_i$, of size at most

$$\frac{\log |\mathcal{F}_{R_i}|}{\epsilon} = \frac{\log(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})^d}{\epsilon} = \frac{d}{\epsilon}\big(2 \log \frac{1}{\epsilon} + \log \log \frac{1}{\epsilon}\big).$$

that hits all sets of $\mathcal{F}|_{R_i}$ containing $\Omega(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$ points. Therefore, R' is an ϵ-net for $(X, \mathcal{F})$. □

QUESTIONS

(solutions)

1. Let $(X, \mathcal{F})$ be a set system with VC-dimension 1.
   (a) Show that for any $R \in \mathcal{F}$, and any two sets $R_1, R_2 \in \mathcal{F}$, either $R_1 \cap R \subseteq R_2 \cap R$ or vice versa.
   (b) Prove that there exists an ϵ-net for $(X, \mathcal{F})$ of size $O(1/\epsilon)$.
2. Let $(X, \mathcal{F})$ be a set system with VC-dimension d, with $|X| = n$.
   (a) Show that for any set $Y \subseteq X$ with $|Y| \geq d + 1$, there exists $Y' \subseteq Y$ such that any set containing $Y \setminus Y'$ contains a point of Y'. We say that Y' *hits* $Y \setminus Y'$.
   (b) Given $Y \subseteq X$, there exists an $1 \leq i \leq d$ and a subset Y' of size i such that Y' hits $\Omega\big(\binom{|Y|}{d+1} / (d\binom{|Y|}{i})\big)$ sets of size $d + 1 - i$ of Y.
   (c) Show that there exists an ϵ-net for $(X, \mathcal{F})$ of size $O(1/\epsilon^d)$ using an idea similar to that of Theorem 4.20.