DCG, Wintersemester 2023/24	Lecture 20 (5 pages)

Set Systems and the VC-Dimension

Anne Driemel

updated: January 13, 2025

In this lecture we will study an important concept which we call a set system. Set systems have wide-ranging applications in learning theory, but we will be mostly concerned with their applications in computational geometry.

1 Definitions and Examples

Definition 20.1 (Set system). *Let X be a set and let $\mathcal{R}$ be a set of subsets of X, that is $\mathcal{R} \subseteq 2^X$ (We denote with 2^X the power set of X). We call $\mathcal{R}$ a set system with ground set X.*

Example 20.2. *The set of all axis-aligned rectangles in the plane defines a set system $\mathcal{R}$ with ground set $X = \mathbb{R}^2$. Formally, each set $r \in \mathcal{R}$ can be specified by a 4-tuple (a, b, c, d) with*

$$r_{a,b,c,d} = \{\ (x, y) \in X \ \ |\ \ a \leq x \leq b, c \leq y \leq d\ \}\,.$$

Example 20.3. *Any finite set system (i.e., for finite X) can be represented as a Boolean matrix with $|X| \times |\mathcal{R}|$ dimensions. An entry at (i, j) of the matrix encodes if the ith element of X is contained in the jth element of $\mathcal{R}$, for some predefined ordering of X and $\mathcal{R}$.*

An important property of a set system is its VC-dimension, named after Vapnik und Chervonenkis. We define it using the following notion of shattering.

Definition 20.4 (Shattering). *We say a set $A \subseteq X$ is shattered by a set system $\mathcal{R}$, if for any set $A' \subseteq A$, there exists a set $r \in \mathcal{R}$ with $A' = r \cap A$. We define the subsystem*

$$\mathcal{R}|_A = \{r \cap A \mid r \in \mathcal{R}\}$$

Note that A is shattered by $\mathcal{R}$ if and only if $\mathcal{R}|_A = 2^A$.

Definition 20.5 (VC-dimension). *The VC-dimension of a set system $\mathcal{R}$ is the maximum cardinality of a set that is shattered by $\mathcal{R}$. We denote it with $\dim(\mathcal{R})$. For the special case $\mathcal{R} = \emptyset$ we define $\dim(\emptyset) = 0$. If there is no maximum cardinality set that is shattered, and if $\mathcal{R} \neq \emptyset$, then we say the VC-dimension is infinite.*

Consider the set system in Example 20.2. The VC-dimension of $\mathcal{R}$ is at least 4, since we can find a 4-element set A of points in the plane that is shattered by $\mathcal{R}$. Figure 1 gives an example of such a set A. On the other hand, no 5-element set can be shattered by $\mathcal{R}$ and this can be seen as follows. Let A be a set of 5 points, $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)$. Consider a sorted ordering of A by x-coordinate, e.g. say

$$x_5 \leq x_3 \leq x_2 \leq x_4 \leq x_1$$

and consider also a sorted ordering of A by y-coordinate, say

$$y_1 \leq y_2 \leq y_5 \leq y_3 \leq y_4$$

Since A contains 5 points, there must be a point $q \in A$ that is neither first nor last in neither list. Note that q is contained in all axis-parallel rectangles that contain the set $A \setminus \{q\}$. Thus

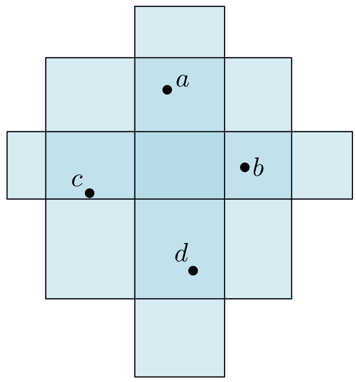

Figure 1: Left: Example set that is shattered by axis-aligned rectangles. For any of the 16 subsets we can find a rectangle that represents the subset by intersection.

A cannot be shattered, since we cannot represent the set $A \setminus \{q\}$. Therefore, the VC-dimension of the set system of axis-parallel rectangles is exactly 4.

As a second example, consider the set system of of all convex polygons in the plane. Its VC-dimension is infinite. Indeed, for any natural number m we can find an m-element set that can be shattered by convex polygons. In particular, let A_m be a set of m points in convex position (The points are in convex position if no point in the set can be written as a convex combination of the others). For any subset $A' \subseteq A$, consider the convex hull of A'. This is a convex polygon that contains the points of A' and none of the points in $A \setminus A'$. Since this construction works for any natural number m, the VC-dimension of this set system is infinite.

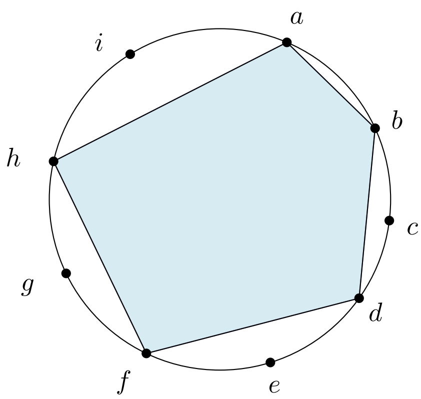

Figure 2: Example of a set of 9 points that is shattered by the set system of convex polygons. The figure shows a convex polygon that represents the subset $\{a, b, d, f, h\}$.

2 Growth of set systems

How many different sets can there be in a set system with $|X| = m$? In general we have $|\mathcal{R}| \leq 2^{|X|} = 2^m$. What if the VC-dimension is very small?

Example 20.6. *Let $X = \{1, 2, \dots, m\}$ and let $\mathcal{R}$ be the set system that contains all subsets of size at most k (for a fixed $k \leq m$). The VC-dimension of this set system is k. We can* *enumerate all sets and see that*

$$\left|\mathcal{R}\right| = \sum_{i=0}^{k} \binom{m}{i} .$$

Note that, in this summation, we have:

$$\binom{m}{i} = \frac{m!}{(m-i)! \cdot i!} \leq \frac{m^i}{i!} = \left(\frac{m}{k}\right)^i \frac{k^i}{i!} \leq \left(\frac{m}{k}\right)^k \frac{k^i}{i!} .$$

Together with the series definition of the exponential function $\mathrm{e}^x = \sum_{i=0}^{\infty}(x^i/i!)$, *we get*

$$\sum_{i=0}^{k} \binom{m}{i} \leq \sum_{i=0}^{k} \left(\frac{m}{k}\right)^k \frac{k^i}{i!} = \left(\frac{m}{k}\right)^k \sum_{i=0}^{k} \frac{k^i}{i!} \leq \left(\frac{m}{k}\right)^k \mathrm{e}^k = \left(\frac{\mathrm{e}}{k}\right)^k m^k .$$

Thus, for any fixed k*, the number of sets grows only polynomially in the size of the set system* m*, instead of exponentially in* m*.*

We now want to show a general upper bound on the cardinality of set systems in terms of the number of elements of the ground set and the VC-dimension. We will see that the above example is maximal in this sense and that it exemplifies a characteristic behaviour.

Lemma 20.7 (Sauer-Shelah)**.** *For any set system* $\mathcal{R}$ *with an* m*-element ground set* X *and VC-dimension* d, *it holds that*

$$\left|\mathcal{R}\right| \leq \sum_{i=0}^{\mathrm{d}} \binom{m}{i} .$$

Proof. We show the statement by induction on m.

As a base case we take $m = 0$. In this case $\mathcal{R}$ can at most contain the empty set, or be the empty set, and thus $\left|\mathcal{R}\right| \leq 1$. At the same time we have by the definition of the binomial coefficient that $\binom{0}{0} = 1$. Since $1 = \sum_{i=0}^{d} \binom{0}{i}$, this shows correctness in the base case.

For the induction step, we consider $m > 0$. Assume first that $\mathrm{d} = 0$. In this case, no single-element subset of X is shattered. This means that each element of X is either contained in all sets $r \in \mathcal{R}$, or in none of them. Thus, all sets $r \in \mathcal{R}$ must be the same, so $\left|\mathcal{R}\right| \leq 1$. Since $1 = \sum_{i=0}^{0} \binom{m}{i}$, this proves correctness also in this case.

Now consider the case $\mathrm{d} > 0$. Let $x \in X$ be fixed and consider the set system

$$\mathcal{R}_1 = \{\ r \setminus \{x\} \ \mid\ r \in \mathcal{R}\ \} .$$

Let its VC-dimension be denoted with d_1. Note that $\mathrm{d}_1 \leq \mathrm{d}$, since any set $A \subseteq X \setminus \{x\}$ that is shattered by $\mathcal{R}_1$ is also shattered by $\mathcal{R}$, so $\mathrm{d} \geq \mathrm{d}_1$.

Now the induction hypothesis implies:

$$\left|\mathcal{R}_1\right| \leq \sum_{i=0}^{\mathrm{d}_1} \binom{m-1}{i} \leq \sum_{i=0}^{\mathrm{d}} \binom{m-1}{i} .$$

This does not immediately give us a good bound on $\left|\mathcal{R}\right|$ though, since for a set $(r \setminus \{x\}) \in \mathcal{R}_1$, there could be two sets in $\mathcal{R}$, namely $r \setminus \{x\}$ and $r \cup \{x\}$, and $2\sum_{i=0}^{\mathrm{d}} \binom{m-1}{i}$ may be more than $\sum_{i=0}^{\mathrm{d}} \binom{m}{i}$ (for example, if $\mathrm{d} = 1$ and $m = 2$). Indeed, if the size of the system would indeed increase by a factor two with every induction step, we would obtain a bound that is exponential instead of polynomial in m.

Therefore we will count the pairs of "twin" sets $r \setminus \{x\}$ and $r \cup \{x\}$ in $\mathcal{R}$ more precisely. To do so, we define a second set system:

$$\mathcal{R}_2 = \{\ r \setminus \{x\}\ \mid\ r \setminus \{x\} \in \mathcal{R} \text{ and } r \cup \{x\} \in \mathcal{R}\ \}.$$

Note that for each set in $\mathcal{R}_2$ there are two sets in $\mathcal{R}$ that collapse into one set when we go from $\mathcal{R}$ to $\mathcal{R}_1$ by restricting the ground set to X and no other sets are destroyed in this process.

Therefore, we can now precisely count the number of sets in $\mathcal{R}$. It holds that

$$\left|\mathcal{R}\right| = \left|\mathcal{R}_1\right| + \left|\mathcal{R}_2\right|. \tag{1}$$

Now, let $\mathrm{d}_2 = \dim(\mathcal{R}_2)$. We claim that $\mathrm{d}_2 \leq \mathrm{d} - 1$.

For the sake of contradiction, assume that the VC dimension of $\mathcal{R}_2$ is at least d. Then there exists a set $A \subseteq X \setminus \{x\}$ with $|A| = \mathrm{d}$ and such that A is shattered by $\mathcal{R}_2$. But then it must be that the set $A \cup \{x\}$, of size $\mathrm{d} + 1$, is shattered by $\mathcal{R}$, since $\mathcal{R}_2$ only contains sets $r \setminus \{x\}$ such that $r \setminus \{x\}$ and $r \cup \{x\}$ are both in $\mathcal{R}$. This would contradict the assumption that the VC-dimension of $\mathcal{R}$ is equal to d.

Therefore, by induction we have that

$$|\mathcal{R}_2| \leq \sum_{i=0}^{\mathrm{d}_2} \binom{m-1}{i} \leq \sum_{i=0}^{\mathrm{d}-1} \binom{m-1}{i} = \sum_{j=1}^{\mathrm{d}} \binom{m-1}{j-1}$$

By substituting into Equation 1 we get

$$\left|\mathcal{R}\right| \leq \sum_{i=0}^{\mathrm{d}} \binom{m-1}{i} + \sum_{j=1}^{\mathrm{d}} \binom{m-1}{j-1} = 1 + \sum_{i=1}^{\mathrm{d}} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) = \sum_{i=0}^{\mathrm{d}} \binom{m}{i},$$

where the last equality follows from the recursive formula of the binomial coefficient. □

Finally, we want to extend and sharpen the statement of Lemma 20.7. In particular, we are interested in the cardinality of the set system when restricted to a finite subset of the ground set. We have already considered subsystems and observed that the VC-dimension cannot increase when going to a subsystem.

Theorem 20.8. *Let $\mathcal{R}$ be a set system with ground set X and VC-dimension $\mathrm{d} \geq 1$. For any natural number m it holds that*

$$\max_{\substack{A \subseteq X \\ |A| = m}} \left|\mathcal{R}|_A\right| \leq \left(\frac{\mathrm{e}m}{\mathrm{d}}\right)^{\mathrm{d}}.$$

We call $\Pi_{\mathcal{R}}$, defined by:

$$\Pi_{\mathcal{R}}(m) := \max_{\substack{A \subseteq X \\ |A| = m}} \left|\mathcal{R}|_A\right|,$$

the growth function or shatter function of $\mathcal{R}$.

Proof. Since VC-dimension cannot increase by restricting to a subsystem, we have for any subsystem defined by a set $A \subseteq X$ that $\dim(\mathcal{R}|_A) \leq d$.

Thus, we can directly apply Lemma 20.7 and get for any such A with $\left|A\right| = m$:

$$\left|\mathcal{R}|_A\right| \leq \sum_{i=0}^{\mathrm{d}} \binom{m}{i}.$$

By the same calculation as in Example 20.6 we can bound this sum to $\left(\frac{\mathrm{e}m}{\mathrm{d}}\right)^{\mathrm{d}}$. Since we show the bound for any m-element set A, it also holds for the maximum over all such sets. This concludes the proof. □

References

- Sariel Har-Peled, Chapter 5 in *Geometric Approximation Algorithms.* AMS Mathematical Surveys and Monographs. 2011.
- Jiří Matoušek, Chapters 10.2 and 10.3 in *Lectures on Discrete Geometry*, Springer Graduate Texts in Mathematics.