Chapter 13

Epsilon Nets

13.1 Motivation

Here is our scenario for this chapter. We are given a set A of points in $\mathbb{R}^d$ and a family $\mathcal{R}$ of *ranges* $r \subseteq \mathbb{R}^d$, for example the set of all balls, halfspaces, or convex sets in $\mathbb{R}^d$. A is huge and probably not even known completely; similarly, $\mathcal{R}$ may not be accessible explicitly (in the examples above, it is an uncountable set). Still, we want to learn something about A and $\mathcal{R}$.

The situation is familiar, definitely, if we don't insist on the geometric setting. For example, let A be the set of consumers living in Switzerland, and let $\tilde{r}$ be the subset of consumers who frequently eat a certain food product, say Lindt chocolate. We have similar subsets for other food products, and together, they form the family of ranges $\mathcal{R}$.

If we want to learn something about $\tilde{r}$, e.g. the ratio $\frac{|\tilde{r}|}{|A|}$ (the fraction of consumers frequently eating Lindt chocolate), then we typically sample a subset S of A and see what portion of S lies in $\tilde{r}$. We want to believe that

$$\frac{|\tilde{r} \cap S|}{|S|} \quad \text{approximates} \quad \frac{|\tilde{r}|}{|A|},$$

and statistics tells us to what extent this is justified. In fact, consumer surveys are based on this approach: in our example, S is a sample of consumers who are being asked about their chocolate preferences. After this, the quantity $|\tilde{r} \cap S|/|S|$ is known and used to predict the "popularity" $|\tilde{r}|/|A|$ of Lindt chocolate among Swiss consumers.

In this chapter, we consider a different kind of approximation. Suppose that we are interested in the most popular food products in Switzerland, the ones which are frequently eaten by more than an ε-fraction of all consumers, for some fixed $0 \leqslant \varepsilon \leqslant 1$. The goal is to find a small subset N of consumers that "represent" all popular products. Formally, we want to find a set $N \subseteq A$ such that

$$\text{for all } r\text{:} \quad \frac{|r|}{|A|} > \varepsilon \quad \Rightarrow \quad r \cap N \neq \emptyset.$$

Such a subset is called an *epsilon net*. Obviously, $N = A$ is an epsilon net for all ε, but as already mentioned above, the point here is to have a *small* set N.

Epsilon nets are very useful in many contexts that we won't discuss here. But already in the food consumption example above, it is clear that a small representative set of consumers is a good thing to have; for example if you quickly need a statement about a particular popular food product, you know that you will find somebody in your representative set who knows the product.

The material of this chapter is classic and goes back to Haussler and Welzl [1].

13.2 Range spaces and ε-nets.

Here is the formal framework. Let X be a (possibly infinite) set and $\mathcal{R} \subseteq 2^X$. The pair $(X, \mathcal{R})$ is called a *range space*[1], with X its *points* and the elements of $\mathcal{R}$ its *ranges*.

Definition 13.1 *Let $(X, \mathcal{R})$ be a range space. Given $A \subseteq X$, finite, and $\varepsilon \in \mathbb{R}$, $0 \leqslant \varepsilon \leqslant 1$, a subset N of A is called an* ε-net of A (w.r.t. $\mathcal{R}$) *if*

$$\text{for all } r \in \mathcal{R}: \quad |r \cap A| > \varepsilon |A| \quad \Rightarrow \quad r \cap N \neq \emptyset \ .$$

This definition is easy to write down, but it is not so easy to grasp, and this is why we will go through a couple of examples below. Note that we have a slightly more general setup here, compared to the motivating Section 13.1 where we had $X = A$.

13.2.1 Examples

Typical examples of range spaces in our geometric context are

- $(\mathbb{R}, \mathcal{H}_1)$ with $\mathcal{H}_1 := \{(-\infty, a] \mid a \in \mathbb{R}\} \cup \{[a, \infty) \mid a \in \mathbb{R}\}$ (*half-infinite intervals*), and
- $(\mathbb{R}, \mathcal{I})$ with $\mathcal{I} := \{[a, b] \| \, a, b \in \mathbb{R}, a \leqslant b\}$ (*intervals*),

and higher-dimensional counter-parts

- $(\mathbb{R}^d, \mathcal{H}_d)$ with $\mathcal{H}_d$ the closed *halfspaces* in $\mathbb{R}^d$ bounded by hyperplanes,
- $(\mathbb{R}^d, \mathcal{B}_d)$ with $\mathcal{B}_d$ the closed *balls* in $\mathbb{R}^d$,
- $(\mathbb{R}^d, \mathcal{S}_d)$ with $\mathcal{S}_d$ the d-dimensional *simplices* in $\mathbb{R}^d$, and
- $(\mathbb{R}^d, \mathcal{C}_d)$ with $\mathcal{C}_d$ the *convex sets* in $\mathbb{R}^d$.

[1] In order to avoid confusion: A range space is nothing else but a set system, sometimes also called hypergraph. It is the context, where we think of X as points and $\mathcal{R}$ as ranges in some geometric ambient space, that suggests the name at hand.

ε-Nets w.r.t. $(\mathbb{R}, \mathcal{H}_1)$ are particularly simple to obtain. For $A \subseteq \mathbb{R}$, $N := \{\min A, \max A\}$ is an ε-net for every ε—it is even a 0-net. That is, there are ε-nets of size 2, independent from $|A|$ and ε.

The situation gets slightly more interesting for the range space $(\mathbb{R}, \mathcal{I})$ with intervals. Given ε and A with elements

$$a_1 < a_2 < \cdots < a_n \ ,$$

we observe that an ε-net must contain at least one element from any contiguous sequence $\{a_i, a_{i+1}, \ldots, a_{i+k-1}\}$ of $k > \varepsilon n$ (i.e. $k \geqslant \lfloor \varepsilon n \rfloor + 1$) elements in A. In fact, this is a necessary and sufficient condition for ε-nets w.r.t. intervals. Hence,

$$\{a_{\lfloor \varepsilon n \rfloor + 1}, a_{2\lfloor \varepsilon n \rfloor + 2}, \ldots\}$$

is an ε-net of size[2] $\left\lfloor \frac{n}{\lfloor \varepsilon n \rfloor + 1} \right\rfloor \leqslant \lceil \frac{1}{\varepsilon} \rceil - 1$. So while the size of the net depends now on ε, it is still independent of $|A|$.

13.2.2 No point in a large range.

Let us start with a simple exercise, showing that large ranges are easy to "catch". Assume that $|r \cap A| > \varepsilon |A|$ for some fixed r and ε, $0 \leqslant \varepsilon \leqslant 1$.

Now consider the set S obtained by drawing s elements uniformly at random from A (with replacement). We write

$$S \sim A^s,$$

indicating that S is chosen uniformly at random from the set A^s of s-element sequences over A.

What is the probability that $S \sim A^s$ fails to intersect with r, i.e. $S \cap r = \emptyset$? For $p := \frac{|r \cap A|}{|A|}$ (note $p > \varepsilon$) we get[3]

$$\text{prob}(S \cap r = \emptyset) = (1-p)^s < (1-\varepsilon)^s \leqslant e^{-\varepsilon s} \ .$$

That is, if $s = \frac{1}{\varepsilon}$ then this probability is at most $e^{-1} \approx 0.368$, and if we choose $s = \lambda \frac{1}{\varepsilon}$, then this probability decreases exponentially with λ: It is at most $e^{-\lambda}$.

For example, if $|A| = 10000$ and $|r \cap A| > 100$ (r contains more than 1% of the points in A), then a sample of 300 points is disjoint from r with probability at most $e^{-3} \approx 0.05$.

---

[2] The number L of elements in the set is the largest ℓ such that $\ell(\lfloor \varepsilon n \rfloor + 1) \leqslant n$, hence $L = \left\lfloor \frac{n}{\lfloor \varepsilon n \rfloor + 1} \right\rfloor$. Since $\lfloor \varepsilon n \rfloor + 1 > \varepsilon n$, we have $\frac{n}{\lfloor \varepsilon n \rfloor + 1} < \frac{1}{\varepsilon}$, and so $L < \frac{1}{\varepsilon}$, i.e. $L \leqslant \lceil \frac{1}{\varepsilon} \rceil - 1$.

[3] We make use of the inequality $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$.

13.2.3 Smallest enclosing balls.

Here is a potential use of this for a geometric problem. Suppose A is a set of n points in $\mathbb{R}^d$, and we want to compute the smallest enclosing ball of A. In fact, we are willing to accept some mistake, in that, for some given ε, we want a small ball that contains all but at most εn points from A. So let's choose a sample S of $\lambda\frac{1}{\varepsilon}$ points drawn uniformly (with replacement) from A and compute the smallest enclosing ball B of S. Now let $r := \mathbb{R}^d \setminus B$, the complement of B in $\mathbb{R}^d$, play the role of the range in the analysis above. Obviously $r \cap S = \emptyset$, so it is unlikely that $|r \cap A| > \varepsilon|A|$, since—if so—the probability of $S \cap r = \emptyset$ was at most $e^{-\lambda}$.

It is important to understand that this was complete nonsense!

For the probabilistic analysis above we have to first choose r and then draw the sample—and not, as done in the smallest ball example, first draw the sample and then choose r *based on the sample*. That cannot possibly work, since we could always choose r simply as the complement $\mathbb{R}^d \setminus S$—then clearly $r \cap S = \emptyset$ and $|r \cap A| > \varepsilon|A|$, unless $|S| \geqslant (1-\varepsilon)|A|$.

While you hopefully agree on this, you might find the counterargument with $r = \mathbb{R}^d \setminus S$ somewhat artificial, e.g. complements of balls cannot be that selective in 'extracting' points from A. It is exactly the purpose of this chapter to understand to what extent this advocated intuition is justified or not.

13.3 Either almost all is needed or a constant suffices.

Let us reveal the spectrum of possibilities right away, although its proof will have to await some preparatory steps.

Theorem 13.2 *Let $(X, \mathcal{R})$ be an infinite range space. Then one of the following two statements holds.*

(1) For every $n \in \mathbb{N}$ there is a set $A_n \subseteq X$ with $|A_n| = n$ such that for every ε, $0 \leqslant \varepsilon \leqslant 1$, an ε-net must have size at least $(1-\varepsilon)n$.

(2) There is a constant δ depending on $(X, \mathcal{R})$, such that for every finite $A \subseteq X$ and every ε, $0 < \varepsilon \leqslant 1$, there is an ε-net of A w.r.t. $\mathcal{R}$ of size at most $\frac{8\delta}{\varepsilon}\log_2\frac{4\delta}{\varepsilon}$ (independent of the size of A).

That is, either we have always ε-nets of size independent of $|A|$, or we have to do the trivial thing, namely choosing all but εn points for an ε-net. Obviously, the range spaces $(\mathbb{R}, \mathcal{H}_1)$ and $(\mathbb{R}, \mathcal{I})$ fall into category (2) of the theorem.

For an example for (1), consider $(\mathbb{R}^2, \mathcal{C}_2)$. For any $n \in \mathbb{N}$, let A_n be a set of n points in convex position. For every $N \subseteq A_n$ there is a range $r \in \mathcal{C}_2$, namely the convex hull of $A_n \setminus N$, such that $A_n \cap r = A_n \setminus N$ (hence, $r \cap N = \emptyset$); see Figure 13.1. Therefore,

$N \subseteq A_n$ cannot be an ε-net of A_n w.r.t. $\mathcal{C}_2$ if $|A_n \setminus N| = n - |N| > \varepsilon n$. Consequently, an ε-net must contain at least $n - \varepsilon n = (1-\varepsilon)n$ points.[4]

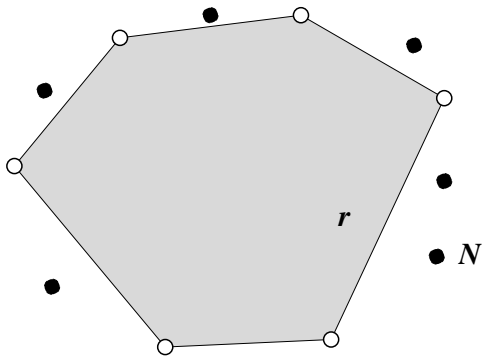

Figure 13.1: *If $\mathcal{R}$ consists of all convex sets in the plane, then only trivial epsilon nets exist: for every subset N (black points) of a set A_n in convex position, the range $r = \mathrm{conv}(A_n \setminus N)$ fails to intersect N.*

So what distinguishes $(\mathbb{R}^2, \mathcal{C}_2)$ from $(\mathbb{R}, \mathcal{H}_1)$ and $(\mathbb{R}, \mathcal{I})$? And which of the two cases applies to the many other range spaces we have listed above? Will all of this eventually tell us something about our attempt of computing a small ball covering all but at most εn out of n given points? This and more should be clear by the end of this chapter.

13.4 What makes the difference: VC-dimension

Given a range space $(X, \mathcal{R})$ and $A \subseteq X$, we let

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

the *projection of $\mathcal{R}$ to A*. Even if $\mathcal{R}$ is infinite, $\mathcal{R}\big|_A$ is always of size at most 2^n if A is an n-element set. The significance of projections in our context becomes clear if we rewrite Definition 13.1 in terms of projections: $N \subseteq A$ is an ε-net if

$$\text{for all } r \in \mathcal{R}\big|_A\colon \quad |r| > \varepsilon|A| \quad \Rightarrow \quad r \cap N \neq \emptyset.$$

All of a sudden, the conditions for an ε-net have become discrete, and they only depend on the finite range space $(A, \mathcal{R}\big|_A)$.

[4] If we were satisfied with any abstract example for category (1), we could have taken $(X, 2^X)$ for any infinite set X.

Note that, for A a set of n points in convex position in the plane, $\mathcal{C}_2\big|_A = 2^A$; we get every subset of A by an intersection with a convex set (this is also the message of Figure 13.1). That is $|\mathcal{C}_2\big|_A| = 2^n$, the highest possible value.

For A a set of n points in $\mathbb{R}$, we can easily see that[5] $|\mathcal{I}\big|_A| = \binom{n+1}{2} + 1 = O(n^2)$. A similar argument shows that $|\mathcal{H}_1\big|_A| = 2n$. Now comes the crucial definition.

Definition 13.3 *Given a range space $(X, \mathcal{R})$, a subset A of X is* shattered by $\mathcal{R}$ *if $\mathcal{R}\big|_A = 2^A$. The* VC-dimension[6] of $(X, \mathcal{R})$, $\mathrm{VCdim}(X, \mathcal{R})$, *is the cardinality (possibly infinite) of the largest subset of X that is shattered by $\mathcal{R}$. If no set is shattered (i.e. not even the empty set which means that $\mathcal{R}$ is empty), we set the VC-dimension to -1.*

We had just convinced ourselves that $(\mathbb{R}^2, \mathcal{C}_2)$ has arbitrarily large sets that can be shattered. Therefore, $\mathrm{VCdim}(\mathbb{R}^2, \mathcal{C}_2) = \infty$.

Consider now $(\mathbb{R}, \mathcal{I})$. Two points $A = \{a, b\}$ can be shattered, since for each of the 4 subsets, $\emptyset$, $\{a\}$, $\{b\}$, and $\{a, b\}$, of A, there is an interval that generates that subset by intersection with A. However, for $A = \{a, b, c\}$ with $a < b < c$ there is no interval that contains a and c but not b. Hence, $\mathrm{VCdim}(\mathbb{R}, \mathcal{I}) = 2$.

Exercise 13.4 *What is $\mathrm{VCdim}(\mathbb{R}, \mathcal{H}_1)$?*

Exercise 13.5 *Prove that if $\mathrm{VCdim}(X, \mathcal{R}) = \infty$, then we are in case (1) of Theorem 13.2, meaning that only trivial epsilon nets always exist.*

13.4.1 The size of projections for finite VC-dimension.

Here is the (for our purposes) most important consequence of finite VC dimension: there are only polynomially many ranges in every projection.

Lemma 13.6 (Sauer's Lemma) *If $(X, \mathcal{R})$ is a range space of finite VC-dimension at most δ, then*

$$|\mathcal{R}\big|_A| \leqslant \Phi_\delta(n) := \sum_{i=0}^{\delta} \binom{n}{i}$$

for all $A \subseteq X$ with $|A| = n$.

[5] Given A as $a_1 < a_2 \cdots < a_n$ we can choose another $n+1$ points b_i, $0 \leqslant i \leqslant n$, such that

$$b_0 < a_1 < b_1 < a_2 < b_2 < \cdots b_{n-1} < a_n < b_n \ .$$

Each nonempty intersection of A with an interval can be uniquely written as $A \cap [b_i, b_j]$ for $0 \leqslant i < j \leqslant n$. This gives $\binom{n+1}{2}$ plus one for the empty set.

[6] 'VC' in honor of the Russian statisticians V. N. Vapnik and A. Ya. Chervonenkis, who discovered the crucial role of this parameter in the late sixties.

Proof. First let us observe that $\Phi : \mathbb{N}_0 \cup \{-1\} \times \mathbb{N}_0 \to \mathbb{N}_0$ is defined by the recurrence[7]

$$\Phi_\delta(n) = \begin{cases} 0 & \delta = -1, \\ 1 & n = 0 \text{ and } \delta \geqslant 0, \text{ and} \\ \Phi_\delta(n-1) + \Phi_{\delta-1}(n-1) & \text{otherwise.} \end{cases}$$

Second, we note that the VC-dimension cannot increase by passing from $(X, \mathcal{R})$ to a projection (A, R), $R := \mathcal{R}\big|_A$. Hence, it suffices to consider the finite range space (A, R)—which is of VC-dimension at most δ—and show $|R| \leqslant \Phi_\delta(n)$ (since Φ is monotone in δ).

Now we proceed to a proof by induction of this inequality. If $A = \emptyset$ or $R = \emptyset$ the statement is trivial. Otherwise, we consider the two 'derived' range spaces for some fixed $x \in A$:

$$(A \setminus \{x\}, R - x), \qquad \text{with } R - x := \{r \setminus \{x\} \,|\, r \in R\}$$

(note $R - x = R\Big|_{A \setminus \{x\}}$) and

$$(A \setminus \{x\}, R^{(x)}), \qquad \text{with } R^{(x)} := \{r \in R \,|\, x \notin r, r \cup \{x\} \in R\}.$$

Observe that the ranges in $R^{(x)}$ are exactly those ranges in $R - x$ that have two preimages under the map

$$R \ni \quad r \quad \mapsto \quad r \setminus \{x\} \quad \in R - x \,,$$

all other ranges have a unique preimage. Consequently,

$$|R| = |R - x| + |R^{(x)}| \,.$$

We have $|R - x| \leqslant \Phi_\delta(n-1)$. If $A' \subseteq A \setminus \{x\}$ is shattered by $R^{(x)}$, then $A' \cup \{x\}$ is shattered by R. Hence, $(A \setminus \{x\}, R^{(x)})$ has VC-dimension at most $\delta - 1$ and $|R^{(x)}| \leqslant \Phi_{\delta-1}(n-1)$. Summing up, it follows that

$$|R| \leqslant \Phi_\delta(n-1) + \Phi_{\delta-1}(n-1) = \Phi_\delta(n)$$

which yields the assertion of the lemma. □

In order to see that the bound given in the lemma is tight, consider the range space

$$(X, \bigcup_{i=0}^{\delta} \binom{X}{i}) \,.$$

Obviously, a set of more than δ elements cannot be shattered (hence, the VC-dimension is at most δ), and for any finite $A \subseteq X$, the projection of the ranges to A is $\bigcup_{i=0}^{\delta} \binom{A}{\delta}$—with cardinality $\Phi_\delta(|A|)$.

[7] We recall that the binomial coefficients $\binom{n}{k}$ (with $k, n \in \mathbb{N}_0$) satisfy the recurrence $\binom{n}{k} = 0$ if $n < k$, $\binom{n}{0} = 1$, and $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.

We note that a rough, but for our purposes good enough estimate for Φ is given by[8]

$$\Phi_\delta(n) \leqslant n^\delta \quad \text{for } \delta \geqslant 2.$$

We have seen now that the maximum possible size of projections either grows exponentially (2^n in case of infinite VC-dimension) or it is bounded by a polynomial n^δ in case of finite VC-dimension δ). The latter is the key to the existence of small ε-nets. Before shedding light on this, let us better understand when the VC-dimension is finite.

13.5 VC-dimension of Geometric Range Spaces

13.5.1 Halfspaces.

Let us investigate the VC-dimension of $(\mathbb{R}^2, \mathcal{H}_2)$. It is easily seen that three points in the plane can be shattered by halfplanes, as long a they do not lie on a common line. Hence, the VC-dimension is at least 3. Now consider 4 points. If three of them lie on a common line, there is no way to separate the middle point on this line from the other two by a halfplane. So let us assume that no three points lie on a line. Either three of them are vertices of a triangle that contains the fourth point—then we cannot possibly separate the fourth point from the remaining three points by a halfplane. Or the four points are vertices of a convex quadrilateral—then there is no way of separating the endpoints from a diagonal from the other two points. Consequently, four points cannot be shattered, and $\mathrm{VCdim}(\mathbb{R}^2, \mathcal{H}_2) = 3$ is established.

The above argument gets tedious in higher dimensions, if it works in a rigorous way at all. Fortunately, we can employ a classic: By Radon's Lemma (Theorem 3.8), every set A of $\geqslant d+2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint subsets A_1 and A_2 such that $\mathrm{conv}(A_1) \cap \mathrm{conv}(A_2) \neq \emptyset$. We get, as an easy implication, that a set A of at least $d+2$ points in $\mathbb{R}^d$ cannot be shattered by halfspaces. Indeed, let $A_1 \mathbin{\dot\cup} A_2$ be a partition as guaranteed by Radon's Lemma. Now every halfspace containing A_1 must contain at least one point of A_2, hence $h \cap A = A_1$ is impossible for a halfspace h and thus A is not shattered by $\mathcal{H}_d$. Moreover, it is easily seen that the vertex set of a d-dimensional simplex (there are $d+1$ vertices) can be shattered by halfspaces (each subset of the vertices forms a face of the simplex and can thus be separated from the rest by a hyperplane). We summarize that

$$\mathrm{VCdim}(\mathbb{R}^d, \mathcal{H}_d) = d+1 \ .$$

Let us now consider the range space $(\mathbb{R}^d, \check{\mathcal{H}}_d)$, where $\check{\mathcal{H}}_d$ denotes the set of all closed halfspaces below non-vertical hyperplanes[9]–we call these *lower halfspaces*. Since $\check{\mathcal{H}}_d \subseteq$

[8] A better estimate, at least for $\delta \geqslant 3$, is given by $\Phi_\delta(n) < \left(\frac{e\,n}{\delta}\right)^d$ for all $n, d \in \mathbb{N}$, $d \leqslant n$.

[9] A hyperplane is called non-vertical if it can be specified by a linear equation $\sum_{i=1}^{d} \lambda_i x_i = \lambda_{d+1}$ with $\lambda_d \neq 0$; see also Section 1.2

$\mathcal{H}_d$, the VC-dimension of $(\mathbb{R}^d, \check{\mathcal{H}}_d)$ is at most $d+1$, but, in fact, it not too difficult to show

$$\mathrm{VCdim}(\mathbb{R}^d, \check{\mathcal{H}}_d) = d \ . \tag{13.7}$$

(Check this claim at least for $d = 2$.) This range space is a geometric example where the bound of Sauer's Lemma is attained. Indeed, for any set A of n points in $\mathbb{R}^d$ in general position[10], it can be shown that

$$|\check{\mathcal{H}}_d|_A| = \Phi_d(n) \ .$$

13.5.2 Balls.

It is easy to convince oneself that the VC-dimension of disks in the plane is 3: Three points not on a line can be shattered and four points cannot. Obviously not, if one of the points is in the convex hull of the other, and for four vertices of a convex quadrilateral, it is not possible for both diagonals to be separated from the endpoints of the respective other diagonal by a circle (if you try to draw a picture of this, you see that you get two circles that intersect four times, which we know is not be the case).

A more rigorous argument which works in all dimensions is looming with the help of (13.7), if we employ the following transformation called *lifting map* that we have already encountered for $d = 2$ in Section 6.3:

$$\begin{array}{rcl} \mathbb{R}^d & \longrightarrow & \mathbb{R}^{d+1} \\ (x_1, x_2, \ldots, x_d) = p & \mapsto & \ell(p) = (x_1, x_2, \ldots, x_d,\ {x_1}^2 + {x_2}^2 + \cdots + {x_d}^2) \end{array}$$

(For a geometric interpretation, this is a vertical projection of $\mathbb{R}^d$ to the unit paraboloid $x_{d+1} = {x_1}^2 + {x_2}^2 + \cdots + {x_d}^2$ in $\mathbb{R}^{d+1}$.) The remarkable property of this transformation is that it maps balls in $\mathbb{R}^d$ to halfspaces in $\mathbb{R}^{d+1}$ in the following sense.

Consider a ball $B_d(c, \rho)$ ($c = (c_1, c_2, \ldots, c_d) \in \mathbb{R}^d$ the center, and $\rho \in \mathbb{R}^+$ the radius). A point $p = (x_1, x_2, \ldots, x_d)$ lies in this ball iff

$$\begin{aligned} & \sum_{i=1}^{d} (x_i - c_i)^2 \leqslant \rho^2 \Leftrightarrow \sum_{i=1}^{d} ({x_i}^2 - 2x_i c_i + {c_i}^2) \leqslant \rho^2 \\ \Leftrightarrow \quad & \left(\sum_{i=1}^{d} (-2c_i) x_i \right) + ({x_1}^2 + {x_2}^2 + \cdots + {x_d}^2) \leqslant \rho^2 - \sum_{i=1}^{d} {c_i}^2 \ ; \end{aligned}$$

this equivalently means that $\ell(p)$ lies below the non-vertical hyperplane (in $\mathbb{R}^{d+1}$)

$$\begin{aligned} & h = h(c, \rho) = \{x \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} h_i x_i = h_{d+2}\} \ \text{ with} \\ & (h_1, h_2, \ldots, h_d, h_{d+1}, h_{d+2}) = \left((-2c_1), (-2c_2), \ldots, (-2c_d), 1, \rho^2 - \sum_{i=1}^{d} {c_i}^2 \right) \ . \end{aligned}$$

[10] No $i+2$ on a common i-flat for $i \in \{1, 2, \ldots, d-1\}$; in particular, no $d+1$ points on a common hyperplane.

It follows that a set $A \subseteq \mathbb{R}^d$ is shattered by $\mathcal{B}_d$ (the set of closed balls in $\mathbb{R}_d$) iff $\ell(A) := \{\ell(p) \,|\, p \in A\}$ is shattered by $\check{\mathcal{H}}_{d+1}$. Assuming (13.7), this readily yields

$$\mathrm{VCdim}(\mathbb{R}^d, \mathcal{B}_d) = d + 1 \ .$$

The lifting map we have employed here is a special case of a more general paradigm called *linearization* which maps non-linear conditions to linear conditions in higher dimensions.

We have clarified the VC-dimension for all examples of range spaces that we have listed in Section 13.2.1, except for the one involving simplices. Before we elaborate on this, let us prove a first bound on the size of ε-nets when the VC-dimension is finite.

13.6 Small ε-Nets, an Easy Warm-up Version

Theorem 13.8 *Let $n, d \in \mathbb{N}$, $d \geqslant 2$, $\varepsilon \in \mathbb{R}^+$. Let $(X, \mathcal{R})$ be a range space of VC-dimension d. If $A \subseteq X$ with $|A| = n$, then there exists an ε-net N of A w.r.t. $\mathcal{R}$ with $|N| \leqslant \lceil \frac{d \ln n}{\varepsilon} \rceil$.*

Proof. We restrict our attention to the finite projected range space (A, R), $R := \mathcal{R}|_A$, for which we know $|R| \leqslant \Phi_d(n) \leqslant n^d$. It suffices to show that there is a set $N \subseteq A$ with $|N| \leqslant \frac{d \ln n}{\varepsilon}$ which contains an element from each $r \in R_\varepsilon := \{r \in R \,|\, |r| > \varepsilon n\}$.

Suppose, for some $s \in \mathbb{N}$ (to be determined), we let $N \sim A^s$. For each $r \in R_\varepsilon$, we know that $\mathrm{prob}(r \cap N = \emptyset) < (1 - \varepsilon)^s \leqslant e^{-\varepsilon s}$. Therefore,

$$\begin{aligned} \mathrm{prob}(N \text{ is not } \varepsilon\text{-net of } A) &= \mathrm{prob}(\exists r \in R_\varepsilon : r \cap N = \emptyset) \\ &= \mathrm{prob}(\bigvee_{r \in R_\varepsilon} (r \cap N = \emptyset)) \\ &\leqslant \sum_{r \in R_\varepsilon} \mathrm{prob}(r \cap N = \emptyset) < |R_\varepsilon| e^{-\varepsilon s} \leqslant n^d e^{-\varepsilon s} \ . \end{aligned}$$

It follows that if s is chosen so that $n^d e^{-\varepsilon s} \leqslant 1$, then $\mathrm{prob}(N \text{ is not } \varepsilon\text{-net of } A) < 1$ and there remains a positive probability for the event that N is an ε-net of A. Now

$$n^d e^{-\varepsilon s} \leqslant 1 \quad \Leftrightarrow \quad n^d \leqslant e^{\varepsilon s} \quad \Leftrightarrow \quad d \ln n \leqslant \varepsilon s.$$

That is, for $s = \lceil \frac{d \ln n}{\varepsilon} \rceil$, the probability of obtaining an ε-net is positive, and therefore an ε-net of that size has to exist.[11] □

If we are willing to invest a little more in the size of the random sample N, then the probability of being an ε-net grows dramatically. More specifically, for $s = \lceil \frac{d \ln n + \lambda}{\varepsilon} \rceil$, we have

$$n^d e^{-\varepsilon s} \leqslant n^d e^{-d \ln n - \lambda} = e^{-\lambda} \ ,$$

[11] This line of argument "If an experiment produces a certain object with positive probability, then it has to exist", as trivial as it is, admittedly needs some time to digest. It is called *The Probabilistic Method*, and was used and developed to an amazing extent by the famous Hungarian mathematician Paul Erdős starting in the thirties.

and, therefore, a sample of that size is an ε-net with probability at least $1 - e^{-\lambda}$.

We realize that we need $\frac{d \ln n}{\varepsilon}$ sample size to compensate for the (at most) n^d subsets of A which we have to hit—it suffices to ensure positive success probability. The extra $\frac{\lambda}{\varepsilon}$ allows us to boost the success probability.

Also note that if A were shattered by $\mathcal{R}$, then $R = 2^A$ and $|R| = 2^n$. Using this bound instead of n^d in the proof above would require us to choose s to be roughly $\frac{n \ln 2}{\varepsilon}$, a useless estimate which even exceeds n unless ε is large (at least $\ln 2 \approx 0.69$).

13.6.1 Smallest enclosing balls, again

It is time to rehabilitate ourselves a bit concerning the suggested procedure for computing a small ball containing all but at most $\varepsilon|A|$ points from an n-point set $A \subseteq \mathbb{R}^d$.

Let $(\mathbb{R}^d, {\mathcal{B}_d}^{\text{compl}})$ be the range space whose ranges consist of all complements of closed balls in $\mathbb{R}^d$. This has the same VC-dimension $d+1$ as $(\mathbb{R}^d, \mathcal{B}_d)$. Indeed, if $A \cap r = A'$ for a ball r, then $A \cap (\mathbb{R}^d \setminus r) = A \setminus A'$, so A is shattered by $\mathcal{B}_d$ if and only if A is shattered by ${\mathcal{B}_d}^{\text{compl}}$.

Hence, if we choose a sample N of size $s = \lceil \frac{(d+1) \ln n + \lambda}{\varepsilon} \rceil$ then, with probability at least $1 - e^{-\lambda}$ this is an ε-net for A w.r.t. ${\mathcal{B}_d}^{\text{compl}}$. Let us quickly recall what this means: whenever the complement of some ball B has empty intersection with N, then this complement contains at most an $\varepsilon|A|$ points of A. As a consequence, the smallest ball enclosing N has at most $\varepsilon|A|$ points of A outside, with probability at least $1 - e^{-\lambda}$.

13.7 Even Smaller ε-Nets

We still need to prove part 2 of Theorem 13.2, the existence of ε-nets whose size is independent of A. For this, we employ the same strategy as in the previous section, i.e. we sample elements from A uniformly at random, with replacement; a refined analysis will show that—compared to the bound of Theorem 13.8—much less elements suffice. Here is the main technical lemma.

Lemma 13.9 *Let $(X, \mathcal{R})$ be a range space of VC-dimension $\delta \geqslant 2$, and let $A \subseteq X$ be finite. If $x \sim A^m$ for $m \geqslant 8/\varepsilon$, then for the set N_x of elements occurring in x, we have*

$$\text{prob}(N_x \text{ is not an } \varepsilon\text{-net for } A \text{ w.r.t. } \mathcal{R}) \leqslant 2\Phi_\delta(2m)2^{-\frac{\varepsilon m}{2}} .$$

Before we prove this, let us derive the bound of Theorem 13.2 from it. We want that

$$2\Phi_\delta(2m)2^{-\frac{\varepsilon m}{2}} < 1,$$

since then we know that an ε-net of size m exists. We have

$$\begin{aligned}
& 2\Phi_\delta(2m)2^{-\frac{\varepsilon m}{2}} < 1 \\
\Leftarrow\ & 2(2m)^\delta < 2^{\frac{\varepsilon m}{2}} \\
\Leftrightarrow\ & 1 + \delta \log_2(2m) < \frac{\varepsilon m}{2} \\
\Leftarrow\ & 2\delta \log_2 m < \frac{\varepsilon m}{2} \\
\Leftrightarrow\ & \frac{4\delta}{\varepsilon} < \frac{m}{\log_2 m}.
\end{aligned}$$

In the second to last implication, we have used $\delta, m \geqslant 2$. Now we claim that the latter inequality is satisfied for $m = 2\frac{4\delta}{\varepsilon}\log_2\frac{4\delta}{\varepsilon}$. To see this, we need that $\frac{m}{\log_2 m} > \alpha$ for $m = 2\alpha\log_2\alpha$ and $\alpha = \frac{4\delta}{\varepsilon}$. We compute

$$\frac{m}{\log_2 m} = \frac{2\alpha\log_2\alpha}{\log_2(2\alpha\log_2\alpha)} = \frac{2\alpha\log_2\alpha}{1+\log_2\alpha+\log_2\log_2\alpha} = \alpha\frac{2\log_2\alpha}{1+\log_2\alpha+\log_2\log_2\alpha} \geqslant \alpha$$

as long as

$$\log_2\alpha \geqslant 1+\log_2\log_2\alpha \Leftrightarrow \log_2\alpha \geqslant \log_2 2\log_2\alpha \Leftrightarrow \alpha \geqslant 2\log_2\alpha,$$

which holds as long as $\alpha \geqslant 2$, and this is satisfied for $\alpha = \frac{4\delta}{\varepsilon}$.

Proof. (Lemma 13.9) By going to the projection $(A, \{_{\mathcal{R}} A)$, we may assume that $X = A$, so we have a finite range space over n elements (see also Section 13.4). Fix $\varepsilon \in \mathbb{R}^+$. For $t \in \mathbb{N}$, a range $r \in \mathcal{R}$ and $x \in A^t$ (a t-vector of elements from A), we define

$$\text{count}(r, x) = |\{i \in [t] : x_i \in r\}|.$$

Now we consider two events over A^{2m}. The first one is the bad event of not getting an ε-net when we choose m elements at random from A (plus another m elements that don't matter). Recall that $R_\varepsilon = \{r \in \mathcal{R} : |r| > \varepsilon n\}$.

$$Q := \{xy \in A^{2m} \mid \exists r \in R_\varepsilon : \text{count}(r, x) = 0\}$$

Thus $\text{prob}(Q) = \text{prob}(N_x \text{ is not } \varepsilon\text{-net})$ which is exactly what we want to bound.

The second auxiliary event looks somewhat weird.

$$J := \{xy \in A^{2m} \mid x \in A^m, y \in A^m, \exists r \in R_\varepsilon : \text{count}(r, x) = 0 \text{ and } \text{count}(r, y) \geqslant \frac{\varepsilon m}{2}\}.$$

This event satisfies $J \subseteq Q$ and contains pairs of sequences x and y with somewhat contradicting properties for some r. While x fails to contain any element from r, y has many elements from r.

Claim 1. $\text{prob}(J) \leqslant \text{prob}(Q) < 2\text{prob}(J)$.

The first inequality is a consequence of $J \subseteq Q$. To prove the second inequality, we show that

$$\frac{\text{prob}(J)}{\text{prob}(Q)} = \frac{\text{prob}(J \cap Q)}{\text{prob}(Q)} = \text{prob}(J \mid Q) \geqslant \frac{1}{2}.$$

So suppose that $xy \in Q$, with "witness" r, meaning that $r \in R_\varepsilon$ and $\text{count}(r, x) = 0$. We show that $xy \in J$ with probability at least $1/2$, for every fixed such x and y chosen randomly from A^m. This entails the claim.

The random variable $\text{count}(r, y)$ is a sum of m independent Bernoulli experiments with success probability $p := |r|/n > \varepsilon$ (thus expectation p and variance $p(1-p)$). Using linearity of expectation and variance (the latter requires independence of the experiments), we get

$$\begin{aligned} \text{E}(\text{count}(r, y)) &= pm, \\ \text{Var}(\text{count}(r, y)) &= p(1-p)m. \end{aligned}$$

Now we use Chebyshew's inequality[12] to bound the probability of the "bad" event that $\text{count}(r, y) < \frac{\varepsilon m}{2}$. We have

$$\begin{aligned} & \text{prob}\left(\text{count}(r, y) < \frac{\varepsilon m}{2}\right) \\ \leqslant\ & \text{prob}\left(\text{count}(r, y) < \frac{pm}{2}\right) \\ \leqslant\ & \text{prob}\left(|\text{count}(r, y) - \text{E}(\text{count}(r, y))| > \frac{pm}{2}\right) \\ =\ & \text{prob}\left(|\text{count}(r, y) - \text{E}(\text{count}(r, y))| > \frac{1}{2(1-p)}\text{Var}(\text{count}(r, y))\right) \\ \leqslant\ & \frac{4(1-p)^2}{p(1-p)m} = \frac{4(1-p)}{pm} \leqslant \frac{4}{pm} < \frac{4}{\varepsilon m} \leqslant \frac{1}{2}, \end{aligned}$$

since $m \geqslant \frac{8}{\varepsilon}$. Hence

$$\text{prob}\left(\text{count}(r, y) \geqslant \frac{\varepsilon m}{2}\right) \geqslant \frac{1}{2},$$

and the claim is proved.

Now the weird event J reveals its significance: we can nicely bound its probability. The idea is this: We enumerate A as $A = \{a_1, a_2, \ldots, a_n\}$ and let the *type* of $z \in A^t$ be the n-sequence

$$\text{type}(z) = (\text{count}(a_1, z), \text{count}(a_2, z), \ldots, \text{count}(a_n, z)).$$

For example, the type of $z = (1, 2, 4, 2, 2, 4)$ w.r.t. $A = \{1, 2, 3, 4\}$ is $\text{type}(z) = (1, 3, 0, 2)$.

Now fix an arbitrary type τ. We will bound the conditional probability $\text{prob}(xy \in J \mid xy \text{ has type } \tau)$, and the value that we get is independent of τ. It follows that the same value also bounds $\text{prob}(J)$.

[12] $\text{prob}(|X - \text{E}(X)| \geqslant k\text{Var}(X)) \leqslant \frac{1}{k^2 \text{Var}(X)}$

Claim 2. $\text{prob}(xy \in J \mid xy \text{ has type } \tau) \leqslant 2\Phi_\delta(2m)2^{-\varepsilon m/2}$.

To analyze the probability in question, we need to sample $z = xy$ uniformly at random from all sequences of type τ. This can be done as follows: take an arbitrary sequence z' of type τ, and then apply a random permutation $\pi \in S_{2m}$ to obtain $z = \pi(z')$, meaning that

$$z_i = z'_{\pi(i)} \text{ for all } i.$$

Why does this work? First of all, π preserves the type, and it is easy to see that all sequences z of type τ can be obtained in this way. Now we simply count the number of permutations that map z' to a fixed z and see that this number only depends on the τ, so it is the same for all z. Indeed, for every element $a_i \in A$, there are $\tau_i!$ many ways of mapping the a_i's in z' to the a_i's in z. The number of permutations that map z' to z is therefore given by

$$\prod_{i=1}^{n} \tau_i!.$$

By these considerations,

$$\text{prob}(xy \in J \mid xy \text{ has type } \tau) = \text{prob}(\pi(z') \in J)\}.$$

To estimate this, we let S be the set of distinct elements in z' (this is also a function of the type τ). Since $(X, \mathcal{R})$ has VC-dimension δ, we know from Lemma 13.6 that $|\mathcal{R}\big|_S| \leqslant \Phi_\delta(2m)$, so at most that many different subsets T of S can be obtained by intersections with ranges $r \in \mathcal{R}$, and in particular with ranges $r \in R_\varepsilon$.

Now we look at some fixed such T, consider a permutation π and write $\pi(z') = xy$. We call T a *witness* for π if no element of x is in T, but at least $\varepsilon m/2$ elements of y are in T. According to this definition,

$$\pi(z') \in J \quad \Leftrightarrow \quad \pi \text{ has some witness } T \subseteq S.$$

By the union bound,

$$\text{prob}(xy \in J \mid xy \text{ has type } \tau) = \text{prob}(\pi(z') \in J) \leqslant \sum_T \text{prob}(T \text{ is a witness for } \pi).$$

Suppose that z' contains $\ell \geqslant \varepsilon m/2$ occurences of elements of T (for smaller ℓ, T cannot be a witness). The probability of T being a witness for a random permutation is then

$$\frac{\binom{m}{\ell}}{\binom{2m}{\ell}} = \frac{m(m-1)\cdots m(m-\ell+1)}{2m(2m-1)\cdots(2m-l+1)} \leqslant 2^{-\ell} \leqslant 2^{-\frac{\varepsilon m}{2}},$$

since among all the $\binom{2m}{\ell}$ equally likely ways in which π distributes the ℓ occurences in z, exactly the $\binom{m}{\ell}$ equally likely ways of putting them into y are good.[13] Summing up over all at most $\Phi_\delta(2m)$ sets T, we get

$$\text{prob}(xy \in J \mid xy \text{ has type } \tau) \leqslant \Phi_\delta(2m)2^{-\frac{\varepsilon m}{2}},$$

[13] This argument can be made more formal by explicitly counting the permutations that map $\{1, 2, \ldots, \ell\}$ to the set $\{1, 2, \ldots, m\}$. We leave this to the reader.

for all τ.

The proof is finished by combining the two claims:

$$\text{prob}(N_x \text{ is not } \varepsilon\text{-net}) = \text{prob}(Q) \leqslant 2\text{prob}(J) \leqslant 2\Phi_\delta(2m)2^{-\frac{\varepsilon m}{2}}.$$

□

Questions

62. *What is a range space?* Give a definition and a few examples.

63. *What is an epsilon net?* Provide a formal definition, and explain in words what it means.

64. *What is the VC dimension of a range space* Give a definition, and compute the VC dimension of a few example range spaces.

65. *Which range spaces always have small epsilon nets (and how small), and which ones don't?* Explain Theorem 13.2.

66. *How can you compute small epsilon nets?* Explain the general idea, and give the analysis behind the weaker bound of Theorem 13.8.

67. *How can you compute smaller epsilon nets?* Sketch the main steps of the proof of the stronger bound of Theorem 13.2.

References

[1] D. Haussler and Emo Welzl, Epsilon-nets and simplex range queries. *Discrete Comput. Geom.*, **2**, (1987), 127–151.