

ECE 695 “Statistical Learning Theory” Lecture Notes

Lecture 4*

Prof. Shai Ben-David

Cornell University
September 24, 2003

1 Introduction and Reminder

In the previous lectures, our set-up was as follows. We had a set X of elements and a distribution P over the set $X \times \{0, 1\}$. The distribution defines the probability (or density) for each element $x \in X$ to “come up”, and the probability for a certain property of x to hold or not to hold. If $(x, 1)$ “comes up,” then x has the property, and if $(x, 0)$ “comes up,” then x does not have the property. The learning problem consists in predicting, given x generated by P , whether or not this property holds. We are given an independent, identically distributed (i.i.d.) sample $S \in (X \times \{0, 1\})^m$ of distribution P that contains m pairs (x_i, y_i) , $i = 1, 2, \dots, m$, where $x_i \in X$ and $y_i \in \{0, 1\}$, and we use S to find a function $h : X \rightarrow \{0, 1\}$ that takes x and predicts y in a new pair (x, y) .

The prediction function h , called a *hypothesis*, must be selected from a given set H of functions, called the *hypothesis class*. Ideally, we should choose $h \in H$ such that it minimizes the *true error* $E^P(h)$ defined as

$$E^P(h) := \mathbf{P}[(x, y) \sim P : h(x) \neq y]$$

In practice, however, all we can do is to choose $h \in H$ that minimizes the *training error* $E^S(h)$ which equals

$$E^S(h) := \frac{1}{m} \sum_{i=1}^m X_i \quad \text{where} \quad X_i := \begin{cases} 1, & h(x_i) \neq y_i \\ 0, & h(x_i) = y_i \end{cases} \quad \text{and} \quad (x_i, y_i)_{i=1}^m = S$$

For consistency of the learning process, it is important to make sure that $E^S(h)$ converges to $E^P(h)$ as the sample size m tends to infinity. Moreover, the convergence must occur simultaneously for all hypotheses $h \in H$; otherwise even for arbitrarily large sample sizes there may be hypotheses for which the training error is small, but the true error is large. In other words, $E^S(h)$ must *uniformly* converge to $E^P(h)$:

$$\forall \varepsilon > 0 : \lim_{m \rightarrow \infty} \mathbf{P} \left[\begin{array}{l} \text{sample } S \\ \text{of size } m \end{array} : \forall h \in H : |E^P(h) - E^S(h)| < \varepsilon \right] = 1$$

In some situations, only uniform *one-sided* convergence is required to prove consistency. This type of convergence ensures that if the training error $E^S(h)$ is small, then the true error $E^P(h)$ is likely

*Scribed and expanded by Alexandre Evfimievski, aevf@cs.cornell.edu

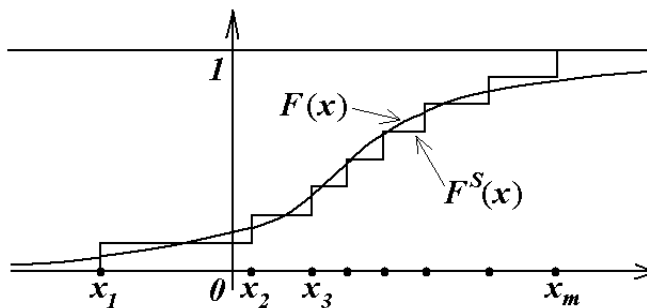


Figure 1: Empirical distribution function.

to be small too, for large enough samples:

$$\forall \varepsilon > 0 : \lim_{m \rightarrow \infty} \mathbf{P} \left[\begin{array}{l} \text{sample } S \\ \text{of size } m \end{array} : \forall h \in H : E^P(h) - E^S(h) < \varepsilon \right] = 1$$

For practical purposes, the convergence should be reasonably fast; it is not enough just to prove it asymptotically. Previously, we considered the case when H is a finite set: $|H| < \infty$. Using the Chernoff-Hoeffding bound, we obtained the formula

$$\forall m, \delta > 0 : \mathbf{P} \left[\forall h \in H : E^P(h) - E^S(h) < \sqrt{\frac{\ln |H| + \ln(1/\delta)}{2m}} \right] > 1 - \delta \quad (1)$$

When the hypothesis class H has infinite cardinality,¹ formula (1) cannot be applied. There are, however, other similar results that provide bounds different from (1).

2 Glivenko-Cantelli Theorem

One result that provides a bound on uniform convergence for an infinite hypothesis class is Glivenko-Cantelli theorem. Its set-up differs from Section 1: Here we want to estimate the distribution function $F(x) = \mathbf{P}[\xi < x]$ of a real-valued random variable ξ from an i.i.d. sample $S = (x_1, x_2, \dots, x_m)$ of size m . The classical approach is to construct an *empirical distribution function* $F^S(x)$ as follows (Fig. 2):

$$F^S(x) := \frac{1}{m} \sum_{i=1}^m \theta(x - x_i), \quad \text{where } \theta(x) := \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

According to the textbook [3], Glivenko-Cantelli theorem, proved in the 1930's long before the advance of statistical learning theory, claims the following:

¹In “naive” mathematics (as opposed to axiomatic set theory) cardinality is just the measure of set size. For a finite set H , cardinality is the number of elements $|H|$. For infinite sets, two sets A and B are defined to be of the same size if there is a one-to-one correspondence between the elements of A and the elements of B . The smallest infinite cardinality is the size of set $\mathbf{N} = \{0, 1, 2, \dots\}$, and it is denoted by \aleph_0 . The sets $\mathbf{N}^2 = \mathbf{N} \times \mathbf{N}$, \mathbf{N}^3 etc. also have size \aleph_0 . After \aleph_0 , the next infinite cardinality is \aleph_1 , then \aleph_2 and so on; they all exist. The set \mathbf{R} of real numbers has cardinality *continuum*, the same as the set of functions $2^{\mathbf{N}} = \{f : \mathbf{N} \rightarrow \{0, 1\}\}$. It is easy to prove that $|2^{\mathbf{N}}| > \aleph_0$; however, the question whether or not $|2^{\mathbf{N}}| = \aleph_1$ cannot be answered within the scope of conventional mathematics.

Theorem 1. (Glivenko-Cantelli) *Empirical distribution function $F^S(x)$ converges to $F(x)$ uniformly for all $x \in \mathbf{R}$:*

$$\forall \varepsilon > 0 : \lim_{m \rightarrow \infty} \mathbf{P} \left[\begin{array}{l} \text{sample } S \\ \text{of size } m \end{array} : \forall x \in \mathbf{R} : |F(x) - F^S(x)| < \varepsilon \right] = 1$$

The convergence in this theorem can be uniformly bounded. At the previous lecture, the bound was given in terms of interval probabilities. Consider a semi-interval $[a, b)$, where $a < b$; we can see that

$$\mathbf{P}[a, b) := \mathbf{P}[x \sim P : x \in [a, b)] = \mathbf{P}[\xi < b] - \mathbf{P}[\xi < a] = F(b) - F(a)$$

and for empirical distribution functions

$$F^S(b) - F^S(a) = \frac{1}{m} \sum_{i=1}^m (\theta(b - x_i) - \theta(a - x_i)) = \frac{|[a, b) \cap S|}{|S|}$$

The uniform bound in the case of intervals states:

$$\forall m, \delta > 0 : \mathbf{P} \left[\forall [a, b) : \left| \mathbf{P}[a, b) - \frac{|[a, b) \cap S|}{|S|} \right| < 2 \sqrt{\frac{2 \log m + \log(1/\delta)}{2(m-2)}} \right] > 1 - \delta \quad (2)$$

We can see that the convergence of the empirical probability estimator to the true probability takes place simultaneously (uniformly) for all intervals $[a, b)$. However, it would be wrong to assume that uniform convergence would take place in other classes of sets. Instead of the class of semi-intervals, consider the class of all finite sets $A = \{a_1, a_2, \dots, a_k\}$ of real numbers. If we fix a finite set, we still have the convergence (by the law of large numbers):

$$\frac{|A \cap S|}{|S|} = \frac{1}{m} \sum_{i=1}^m X_i \rightarrow \mathbf{P}[A] \quad (\text{as } m \rightarrow \infty) \quad \text{where } X_i := \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{otherwise} \end{cases}$$

But if we *do not* fix the finite set and attempt to prove the simultaneous convergence for all finite sets, then we encounter a problem. Even for the uniform probability distribution on the segment $[0, 1]$ we have a contradiction:

$$\forall m \in \mathbf{N} : \mathbf{P} \left[\begin{array}{l} \text{sample } S \\ \text{of size } m \end{array} : \forall A \quad \left| \mathbf{P}[A] - \frac{|A \cap S|}{|S|} \right| < 1 \right] = 0$$

Indeed, for every sample S of size m there always exists a finite set A such that

$$\left| \mathbf{P}[A] - \frac{|A \cap S|}{|S|} \right| = 1,$$

and this set A is the sample itself: $A = S$. We have $\mathbf{P}[A] = 0$ (for every finite set) and the fraction $|A \cap S| / |S| = 1$. So, the class of all finite sets does not allow for uniform convergence of empirical probability estimations.

The question arises: What makes one infinite class of sets have a uniformly converging probability estimator, whereas the other class' estimator does not converge uniformly? The cardinality of both the class of intervals and the class of finite sets is the same: continuum; therefore, the difference lies elsewhere. The true answer to this question, as well as the generalized version of bound (1) for infinite hypothesis sets, was found with the invention of VC-dimension, the core of the modern statistical learning theory.

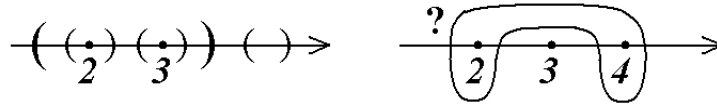


Figure 2: Two sets from Example 1.

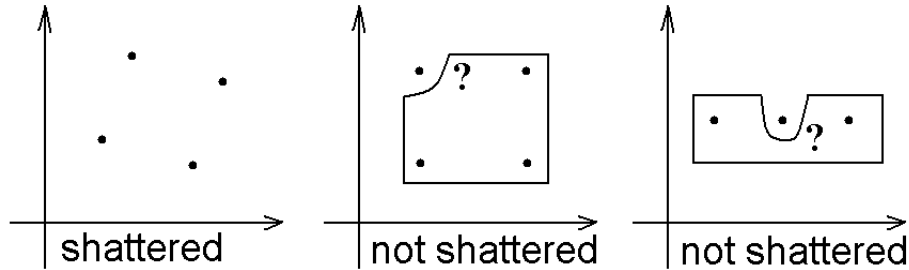


Figure 3: Three situations in Example 2.

3 Definition Of VC-dimension

The notion of VC-dimension was introduced by Vladimir Vapnik and Alexey Chervonenkis [5]; similar notions were studied by Saharon Shelah [2] and Norbert Sauer [1]. Before giving its definition, let us define one combinatorial property of sets:

Definition 1. A collection of sets $H \subseteq 2^X := \{S : S \subseteq X\}$ *shatters* a set $A \subseteq X$ if

$$\{h \cap A : h \in H\} = 2^A := \{B : B \subseteq A\}.$$

In other words, H shatters A if any subset of A can be obtained by intersecting A with some set from collection H .

Consider two simple examples:

Example 1. Let $X = \mathbf{R}$ and $H = \{(a, b) : a < b\}$. Consider two sets: $A = \{2, 3\}$ and $A' = \{2, 3, 4\}$. It is easy to see that the class H of all intervals shatters A and does not shatter A' (Fig. 3). Indeed, we can obtain sets \emptyset , $\{2\}$, $\{3\}$, and $\{2, 3\}$ by intersecting A with intervals. However, for set A' , there is no interval that contains 2 and 4 and does not contain 3; therefore, subset $\{2, 4\}$ cannot be obtained, and A' is not shattered by intervals.

Example 2. Let $X = \mathbf{R}^2$ and

$$H = \{ \text{Rect}(a_1, a_2, b_1, b_2) = \{ \langle x_1, x_2 \rangle : a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2 \} \},$$

that is, H is the collection of axis-aligned rectangles in \mathbf{R}^2 . You can see three situations, one with a set that can be shattered and two with a set that cannot be shattered, on Fig. 3.

Now comes the definition of VC-dimension:

Definition 2. The *VC-dimension* of a collection H of sets is the cardinality of the largest set that is shattered by H . If no such largest set exists, then H has infinite VC-dimension:

$$\text{VC dim}(H) := \max \{ |A| : A \text{ is shattered by } H \}$$

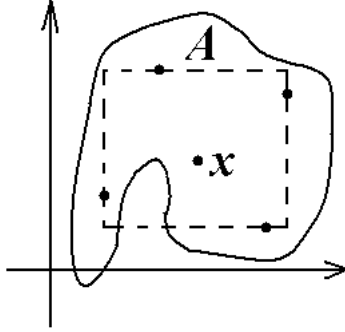


Figure 4: The set A and its five-point subset C from Example 4.

Let us apply this definition to the two previous examples:

Example 3. Let $X = \mathbf{R}$ and $H = \{(a, b) : a < b\}$; then $\text{VC dim}(H) = 2$. Indeed, as we saw in Example 1, there is a set of size 2 that is shattered. Now, take any set A of 3 or more points; then A contains three points $x_1 < x_2 < x_3$. The set

$$B = A \cap ((-\infty, x_1] \cup [x_3, \infty))$$

cannot be obtained by intersecting A with an interval, because any interval that contains x_1 and x_3 must also contain x_2 , and $x_2 \notin B$.

Example 4. Let $X = \mathbf{R}^2$ and let H be the collection of axis-aligned rectangles in \mathbf{R}^2 ; then $\text{VC dim}(H) = 4$. We saw in Example 2 that $\text{VC dim}(H) \geq 4$. Consider any set $A \subset \mathbf{R}^2$ of size 5 or more, and take a five-point subset $C \subseteq A$. In C , take a leftmost point (whose first coordinate is the smallest in C), a rightmost point (first coordinate is the largest), a lowest point (second coordinate is the smallest), and a highest point (second coordinate is the largest); let $x \in C$ be the (fifth) point that was not selected. Now, define $B = A \setminus \{x\}$. It is impossible to make B by intersecting A with an axis-aligned rectangle. Indeed, such a rectangle must contain all four selected points in C ; but in this case the rectangle contains x as well, because its coordinates are within the intervals spanned by selected points (Fig. 3). So, A is not shattered by H , and therefore $\text{VC dim}(H) = 4$.

Example 5. Let $X = \mathbf{R}$ and H be the collection of all finite subsets of \mathbf{R} . Then $\text{VC dim}(H) = \infty$, because any finite set can be shattered by H .

4 VC-dimension Of Half-spaces

Let $X = \mathbf{R}^n$ and $H = HS^n$, where

$$HS^n := \{h(a_1, \dots, a_n, a_{n+1}) : a_i \in \mathbf{R}, i = 1 \dots n + 1\};$$

$$h(a_1, \dots, a_n, a_{n+1}) := \left\{ \langle x_1, \dots, x_n \rangle \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i + a_{n+1} > 0 \right\}$$

In other words, HS^n is the collection of all linear half-spaces in \mathbf{R}^n . This collection occurs frequently in the applications of statistical learning theory. So, let us determine the VC-dimension of HS^n . Before we can do this, we need to give one definition and prove one lemma.

Definition 3. A set $C \subseteq \mathbf{R}^n$ is *convex* if for every pair of points (vectors) $x \in C$, $y \in C$ and for any $\lambda \in [0, 1]$ we have: $\lambda x + (1 - \lambda)y \in C$. Given any set $A \subseteq \mathbf{R}^n$, the *convex hull* of A (denoted $\text{hull}(A)$) is the intersection of all convex sets that contain A :

$$\text{hull}(A) := \bigcap \{C \subseteq \mathbf{R}^n : C \text{ is convex and } A \subseteq C\}$$

Any intersection of convex sets is also a convex set, so $\text{hull}(A)$ is the smallest convex set that contains A . For any finite subset $\{a_1, a_2, \dots, a_k\} \subseteq A$, we have

$$\forall i = 1, \dots, k : \lambda_i \geq 0 \quad \text{and} \quad \sum_{i=1}^k \lambda_i = 1 \quad \implies \quad x = \sum_{i=1}^k \lambda_i a_i \in \text{hull}(A)$$

because x belongs to any convex set that contains points a_1, a_2, \dots, a_k . Note that all half-spaces $h \in HS^n$, as well as their complements $\bar{h} = \mathbf{R}^n \setminus h$, are convex sets.

Lemma 1. (Radon) *For every set $A \subseteq \mathbf{R}^n$, if $|A| \geq n + 2$, then there is a subset $B \subseteq A$ such that*

$$\text{hull}(B) \cap \text{hull}(A \setminus B) \neq \emptyset.$$

*Proof.*² We can assume that $|A| = n + 2$, because if some $A' \subset A$ satisfies the lemma, then A also does. Let $A = \{a_1, a_2, \dots, a_{n+2}\}$; each point $a_i \in A$ has coordinates $a_i = \langle a_{i1}, a_{i2}, \dots, a_{in} \rangle$. Consider the following matrix:

$$M = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & 1 \\ a_{21} & a_{22} & \dots & a_{2n} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n+21} & a_{n+22} & \dots & a_{n+2n} & 1 \end{pmatrix}$$

Matrix M has $n + 1$ columns and $n + 2$ rows, therefore its rows are linearly dependent, which means that there is a linear combination

$$\sum_{i=1}^{n+2} \lambda_i \langle a_{i1}, a_{i2}, \dots, a_{in}, 1 \rangle = \langle 0, \dots, 0 \rangle$$

in which some of λ_i are nonzero. As a consequence, we have

$$\sum_{i=1}^{n+2} \lambda_i a_i = \langle 0, \dots, 0 \rangle \quad \text{and} \quad \sum_{i=1}^{n+2} \lambda_i = 0.$$

Now, let us define $B \subseteq A$ as

$$B = \{a_i \in A : \lambda_i > 0\}$$

It is nonempty, because otherwise all $\lambda_i = 0$. The subset $A \setminus B$ is also nonempty and consists of all $a_i \in A$ such that $\lambda_i \leq 0$. Without loss of generality, assume that the first k points a_1, \dots, a_k are in B , and the rest are in $A \setminus B$. We have:

$$\sum_{i=1}^k \lambda_i a_i = \sum_{i=k+1}^{n+2} (-\lambda_i) a_i \quad \text{and} \quad \sum_{i=1}^k \lambda_i = \sum_{i=k+1}^{n+2} (-\lambda_i) = \Lambda$$

²Taken from website <http://planetmath.org/encyclopedia/RadonsLemma.html>

Denote $\mu_i = \lambda_i/\Lambda$ for $1 \leq i \leq k$ and $\mu_i = (-\lambda_i)/\Lambda$ for $k+1 \leq i \leq n+2$; then all $\mu_i \geq 0$, and we have

$$\sum_{i=1}^k \mu_i a_i = \sum_{i=k+1}^{n+2} \mu_i a_i = x \quad \text{and} \quad \sum_{i=1}^k \mu_i = \sum_{i=k+1}^{n+2} \mu_i = 1$$

This implies that point x belongs to both $\text{hull}(B)$ and $\text{hull}(A \setminus B)$, which proves the lemma. \square

Now we can prove the theorem about the VC-dimension of HS^n .

Theorem 2.

$$\text{VC dim}(HS^n) = n + 1.$$

Proof. Here we shall prove that $\text{VC dim}(HS^n) \leq n + 1$; the proof of $\text{VC dim}(HS^n) \geq n + 1$ is left as an exercise³. Suppose that there exists A such that $|A| \geq n + 2$ and A is shattered by HS^n . Pick B as in Radon's lemma; since A is shattered, there is a half-space $h \in HS^n$ such that $h \cap A = B$. Of course, it is also true that $\bar{h} \cap A = A \setminus B$. By Radon's lemma, there is a point $x \in \text{hull}(B) \cap \text{hull}(A \setminus B)$. The half-space h is a convex set containing B , so $x \in \text{hull}(B) \subseteq h$. The complement \bar{h} is also a convex set, and \bar{h} contains $A \setminus B$, so $x \in \text{hull}(A \setminus B) \subseteq \bar{h}$. Point x belongs to both h and \bar{h} , which is a contradiction. Therefore, A cannot exist, and the theorem is proven. \square

5 The Shatter Function

The shatter function is another combinatorial notion which is closely related to the notion of VC-dimension. This function maps natural numbers to natural numbers, and it measures the "shattering ability" of a given collection of sets.

Definition 4. Given X and the collection $H \subseteq 2^X$, the *shatter function* of H (denoted $\tau_H : \mathbf{N} \rightarrow \mathbf{N}$) is defined as

$$\tau_H(m) := \max_{|A|=m} |\{h \cap A : h \in H\}|$$

In other words, $\tau_H(m)$ is the largest number of subsets one can get from an m -element set by intersecting it with sets from H .

If there is a set of size m that is shattered by H , then $\text{VC dim}(H) \geq m$ and $\tau_H(m) = 2^m$. If there is no such set, then $\text{VC dim}(H) < m$ and $\tau_H(m) < 2^m$.⁴

It turns out that for $m > \text{VC dim}(H)$ not only $\tau_H(m) < 2^m$, but in fact $\tau_H(m) = \text{poly}(m)$. This result is formulated in the following lemma:

Lemma 2. (Sauer) *If H has a finite $\text{VC dim}(H) = d$, then $\forall m \in \mathbf{N} : \tau_H(m) \leq m^d$.*

Proof. We are going to prove even a stronger statement, namely: For any set A and a collection of sets H , we have

$$|\{h \cap A : h \in H\}| \leq |\{B \subseteq A : H \text{ shatters } B\}| \tag{3}$$

This implies Sauer's lemma, because if $\text{VC dim}(H) = d$ then

$$|\{B \subseteq A : H \text{ shatters } B\}| \leq \sum_{i=0}^d \binom{|A|}{i} \leq |A|^d.$$

³Hint: prove that HS^n shatters set $\{\vec{0}, e_1, \dots, e_n\}$ where $\vec{0} = \langle 0, \dots, 0 \rangle$ and $e_i = \langle 0, \dots, 0, 1_i, 0, \dots, 0 \rangle$.

⁴Of course, if there is no set of size m that is shattered, then there is no set of size greater than m that is shattered: all subsets of a shattered set are themselves shattered.

Let us prove inequality (3) by induction. For $A = \emptyset$, both parts of the inequality are equal to 1. Now, let $A \neq \emptyset$; assume that (3) is true for all sets smaller than A and all collections of sets. Pick $x \in A$ and let $A' = A \setminus \{x\}$. Given H , define three new collections of sets:

$$\begin{aligned} H_A &= \{h \cap A : h \in H\}; & H_{A'} &= \{h \cap A' : h \in H\}; \\ H_{A'}^x &= \{B \subseteq A' : B \in H_A \text{ and } B \cup \{x\} \in H_A\} \end{aligned}$$

For every subset $B \subseteq A'$, there are four possibilities:

1. $B \in H_A$ and $B \cup \{x\} \in H_A$; then $B \in H_{A'}$ and gives two counts for $|H_A|$.
2. $B \in H_A$, but $B \cup \{x\} \notin H_A$; then $B \in H_{A'}$ and gives one count for $|H_A|$.
3. $B \notin H_A$, but $B \cup \{x\} \in H_A$; then $B \in H_{A'}$ and gives one count for $|H_A|$.
4. $B \notin H_A$ and $B \cup \{x\} \notin H_A$; then $B \notin H_{A'}$ and gives zero counts for $|H_A|$.

In the first case, B belongs to both $H_{A'}$ and $H_{A'}^x$; in the second and third cases, B belongs only to $H_{A'}$, but not to $H_{A'}^x$. Therefore, we have:

$$|H_A| = |H_{A'}| + |H_{A'}^x| \tag{4}$$

Now we apply the induction hypothesis to A' and two collections $H_{A'}$ and $H_{A'}^x$. Note that both collections contain only subsets of A' , making $h \cap A' = h$. We obtain:

$$\begin{aligned} |H_{A'}| &= |\{h \cap A' : h \in H_{A'}\}| \leq |\{B \subseteq A' : H_{A'} \text{ shatters } B\}| = \\ &= |\{B \subseteq A \setminus \{x\} : H \text{ shatters } B\}| = |\{B \subseteq A : x \notin B \text{ and } H \text{ shatters } B\}| \\ |H_{A'}^x| &= |\{h \cap A' : h \in H_{A'}^x\}| \leq |\{B \subseteq A' : H_{A'}^x \text{ shatters } B\}| = \\ &= |\{B \subseteq A \setminus \{x\} : H \text{ shatters } B \cup \{x\}\}| = |\{B \subseteq A : x \in B \text{ and } H \text{ shatters } B\}| \end{aligned}$$

Combining this with (4), we finally obtain

$$|H_A| = |H_{A'}| + |H_{A'}^x| \leq |\{B \subseteq A : H \text{ shatters } B\}|,$$

which completes the induction. □

Example 6. Suppose $A \subset \mathbf{R}^2$, $|A| = 10\,000$, and $H = HS^2$. By Sauer's lemma, there are at most $10\,000^3 = 10^{12}$ linearly separable subsets of A , out of $2^{10\,000} \approx 10^{3\,000}$ possible subsets.

References

- [1] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.
- [2] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.
- [3] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., 1998.
- [4] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer-Verlag, second edition, 2000.
- [5] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.