Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 6
Scribed by Mac Schwager — T 2/27

In this lecture we complete the elements from VC theory which we set forth in the past lectures. In particular, we prove that the behavior of the growth function is linear until the VC dimension, and logarithmic beyond. This is the property that establishes finite VC dimension as sufficient for the uniform convergence of empirical risk. We also prove the equivalence between the two definitions of VC dimension, in terms of the growth function and shattering. We conclude with examples of various model classes.

Before we proceed, recall the following from last lecture:

- The set of all achievable loss ℓ-tuples given data $z_1, \cdots, z_l$ is:

$$B^\Lambda(z_1, \cdots, z_l) = \{(L(z_1, \alpha), \cdots, L(z_\ell, \alpha)) : \alpha \in \Lambda\},$$

  itself a subset of all possible zero-one ℓ-tuples, i.e. $\{0,1\}^\ell$. The cardinality of $B^\Lambda(z^\ell)$ is denoted by $N^\Lambda(z_1, \cdots, z_l = |B^\Lambda(z_1, \cdots, z_l)|$. In the proof of the main theorem, we have seen how $N^\Lambda(z_1, \cdots, z_l)$ is the number of equivalence classes that produce the same classfication error on the sample set.

- Define the growth function as

$$G^\Lambda(\ell) = \max_{z_1, \cdots, z_l} \ln N^\Lambda(z_1, \cdots, z_l).$$

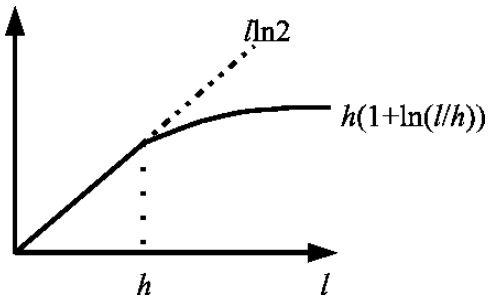
  It serves as a worst case bound on the annealed entropy,

$$H^\Lambda_{ann}(\ell) = \ln \mathbf{E} N^\Lambda(z_1, \cdots, z_l) \leq G^\Lambda(\ell).$$

- The uniform convergence of empirical means property (UCEP) is satisfied if

$$\lim_{\ell \to \infty} \frac{G^\Lambda(\ell)}{\ell} = 0.$$

1 Behavior of the Growth Function



Theorem 1. *The behavior of the growth function, illustrated above, is either:*

$$G^\Lambda(\ell) = \ell \ln 2, \qquad \forall \ell,$$

or there exists $h \in \mathbb{Z}^+$, called the VC dimension of the class, such that:

$$G^\Lambda(\ell) \quad \begin{cases} = & \ell \ln 2 & \text{for} \quad \ell \leq h, \\ \leq & \ln \sum_{i=0}^{h} \binom{\ell}{i} \leq \left(\frac{el}{h}\right)^h = h(1 + \ln \frac{\ell}{h}) & \text{for} \quad \ell > h. \end{cases}$$

Proof. See lecture handout, in the appendix. □

2 VC Dimension

We can rewrite the definition of the VC Dimension of the set of models parameterized by Λ as follows:

$$h = \max\left\{\ell : G^{\Lambda}(\ell) = \ell \ln 2\right\}.$$

If the max does not exist then, by convention, $h = \infty$.

The following corollary immediately follows from this definition and the behavior of the growth function:

Corollary. *If the model class has finite VC dimension, then UCEP holds.*

The alternative definition of VC dimension requires the notion of shattering.

Definition 1 (Shattering)**.** Let $\{A_\alpha\}$ be a family of sets. Consider a given sample set z^ℓ, and let $B = \{z_1, \cdots, z_\ell\}$. We say that $\{A_\alpha\}$ shatters B if $\forall S \subset B\ \exists \alpha \in \Lambda$ such that $S = B \cap A_\alpha$. In other words, $\{A_\alpha\}$ shatters B if it can 'select' any subset of points.

In our setting, we consider a model class $\mathcal{C}$ parametrized by $\alpha \in \Lambda$, and the corresponding indicator loss functions $L(\cdot;\alpha)$. $\mathcal{C}$ generates the family of sets as follows: $A_\alpha = \{z | L(z;\alpha) = 1\}$.

The alternative definition of VC dimension is:

$$h' = \max\{|B| : B \text{ shattered by } \{A_\alpha\}\}.$$

Claim. *We have $h = h'$.*

Proof. We prove the two sides of the equality.

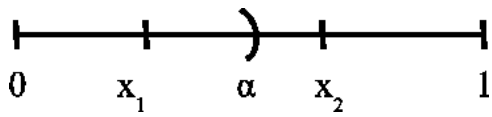
- $h \le h'$ By the definition of h, we know that $G^{\Lambda}(h) = h \ln 2$. By the definition of G^{Λ}, it follows that there exists some z^h such that $N^{\Lambda}(z^h) = 2^h$. This means that all loss h-tuples $(L_1, \cdots, L_h)$ are possible for z^h, and thus each is achievable by some value of α. Let $B = \{z_1, \cdots, z_h\}$, and choose any subset $S \subset B$. Select $\alpha^\star$ that gives a loss h-tuple which is 1 exactly on S. By the definition of A_α, it follows that $A_{\alpha^\star}$ has all the elements of S but none of $B \setminus S$. Therefore $B \cap A_{\alpha^\star} = S$ and, since S was arbitrary, $\{A_\alpha\}$ shatters B. By the definition of h' we then have $h' \ge |B| = h$.
- $h \ge h'$ By the definition of h', we know that there exists some $B = \{z_1, \cdots, z_{h'}\}$ that is shattered by $\{A_\alpha\}$. Choose any loss h'-tuple $(L_1, \cdots, L_{h'})$, and let S be the subset of B where the losses are 1. Due to shatterability, there exists some $\alpha^\star$ such that $B \cap A_{\alpha^\star} = S$. $A_{\alpha^\star}$ has all the elements of S but none of $B \setminus S$, and thus $L(z;\alpha)$ is 1 exactly on S: $(L(z_1;\alpha^\star), \cdots, L(z_{h'};\alpha^\star)) = (L_1, \cdots, L_{h'})$. Since the choice of the loss h'-tuple was arbitrary, it follows that all are achievable, and thus $N^{\Lambda}(z^{h'}) = 2^{h'}$. By the definition of $G^{\Lambda}(h')$, $G^{\Lambda}(h') \ge \ln N^{\Lambda}(z^{h'}) = h' \ln 2$. Since $G^{\Lambda}(\ell) \le \ell \ln 2$, it follows that in fact $G^{\Lambda}(h') = h' \ln 2$. Thus, by the definition of h, we have that $h \ge h'$.

□

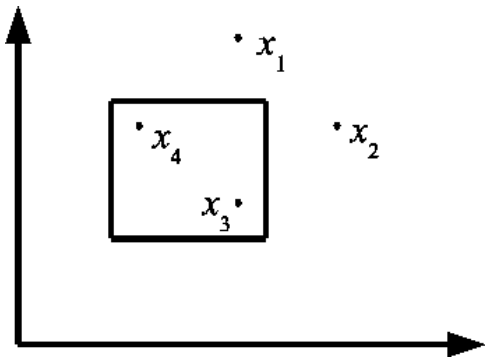
3 Model Class Examples

3.1 Glivenko-Cantelli

$A_\alpha = [0, \alpha)$, $\alpha \in (0, 1]$



We cannot shatter any set of two numbers because we cannot have $j \in A_\alpha$ and $i \notin A_\alpha$ if $j \le i$. VC Dimension is 1.



3.2 Rectangles

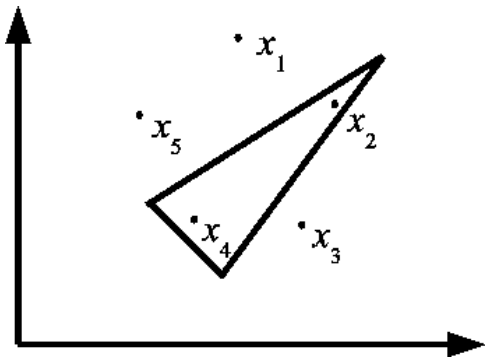
A_α = "All rectangles with sides parallel to the coordinate axes."

VC dimension is at least 4 since we can shatter a set of four points in $\mathbf{R}^2$ arranged in a dia mond. It is claimed that the VC dimension is, in fact, 4.

3.3 Convex Polytopes

A_α = "All convex polytopes in $\mathbf{R}^2$."

Take a set of any number of points in $\mathbf{R}^2$ and arrange them as the vertices of a regular polygon. A_α can shatter any number of points in such an arrangement, so $h = \infty$.



3.4 Separating Hyperplanes

$A_\alpha = \{x | \text{sgn}(b^T x) > 0, \alpha = b \in \mathbf{R}^n\}$

This is the set of separating hyperplanes passing through the origin (the analysis can readily be extended to arbitrary hyperplanes). In $\mathbf{R}^2$, $h = 2$ by inspection.

In $\mathbf{R}^n$ we know the VC dimension is at least n since we can shatter the sample set produced by placing one point on each coordinate axis: $z^n = (z_1, \cdots, z_n)$, where

$$z_i = e_i = (\underset{1}{0}, \cdots, 0, \underset{i}{1}, 0, \cdots, \underset{n}{0}).$$

Claim. *We have: h=n.*

Proof. We already showed that $h \geq n$. We simply need to prove that $h \leq n$. For the sake of contradiction, suppose $h \geq n+1$, so that some sample set of $n+1$ points, which we can represent as a row-concatenated matrix $\mathbf{x} = [x_1 \cdots x_{n+1}] \in \mathbf{R}^{n \times (n+1)}$, is shattered by this class. This means that any possible sign configuration can be achieved by some choice of weight vector b. Let the matrix $\mathbf{B} = [b_1 \cdots b_{2^{n+1}}]$ be the row concatenation of weight vectors generating all 2^{n+1}

distinct sign combinations. Define the matrix $\tilde{V} \in \mathbf{R}^{(n+1)\times 2^{n+1}}$ to be $\tilde{V} = \mathbf{x}'\mathbf{B}$. It follows that $V = \text{sgn}(\tilde{V})$ reproduces all sign configurations in 2^{n+1} columns:

$$V = \underbrace{\begin{bmatrix} + & \cdots & - \\ \vdots & \ddots & \vdots \\ + & \cdots & - \end{bmatrix}}_{2^{n+1}} = \text{sgn}\left(\underbrace{\begin{bmatrix} x_1' \\ \vdots \\ x_{n+1}' \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_{n+1} \end{bmatrix}}_{\tilde{V}} \right).$$

Now, choose an arbitrary vector $c \in \mathbf{R}^{n+1}$ such that $c'\tilde{V} = 0$. Since one of the columns of $\tilde{V}$, say $\tilde{V}_j$, has the same sign configuration as c, it follows that $0 = c'\tilde{V}_j = \sum_{i=1}^{n+1} |c_i||\tilde{V}_j|$, and thus it must be that $c = 0$. This means that $\tilde{V}$ is full rank $n+1$, which cannot be, since it is a product of matrices each of rank less than or equal to n. This is the desired contradiction. □

3.5 Remarks

- In general, the VC dimension is not equal to the number of free parameters in general (refer to the homework for further).
- The VC dimension represents a worst-case scenario. Given a specific distribution, we can sometimes get UCEP even though the VC dimension of the set of functions is ∞.

Proof of the Growth Function Behavior

The proof follows from three lemmas.

Lemma 1. *If for some sample set z^ℓ and for some n we have:*

$$N^\Lambda(z^\ell) > \sum_{i=0}^{n-1} \left(\begin{array}{c} \ell \\ i \end{array} \right),$$

then there exists a subset $z^{\star n}$ of length n, $\{z_1^\star, \cdots, z_n^\star\} \subset \{z_1, \cdots, z_\ell\}$ such that:

$$N^\Lambda(z^{\star n}) = 2^n.$$

Equivalently, the VC dimension of the model class is greater than n.

Proof. Let $\Phi(n, \ell) = \sum_{i=0}^{n-1} \left(\begin{array}{c} \ell \\ i \end{array} \right)$. Φ has the following properties:

$$\Phi(1, \ell) = 1 \qquad \text{(by definition)},$$

$$\Phi(n, \ell) = 2^\ell \qquad \text{if } \ell \leq n+1,$$

$$\Phi(n, 1) = \Phi(n, \ell - 1) + \Phi(n-1, \ell - 1) \qquad \text{if } n \geq 2.$$

The latter follows from the equality:

$$\left(\begin{array}{c} \ell \\ i \end{array} \right) = \left(\begin{array}{c} \ell - 1 \\ i \end{array} \right) + \left(\begin{array}{c} \ell - 1 \\ i - 1 \end{array} \right),$$

summing both sides from $i = 1$ to $n - 1$, we have:

$$\begin{array}{ccccc} \underbrace{\sum_{i=1}^{n-1} \left(\begin{array}{c} \ell \\ i \end{array} \right)} & = & \underbrace{\sum_{i=1}^{n-1} \left(\begin{array}{c} \ell - 1 \\ i \end{array} \right)} & + & \underbrace{\sum_{i=1}^{n-1} \left(\begin{array}{c} ell - 1 \\ i - 1 \end{array} \right)} \\ \Phi(n, 1) - 1 & = & \Phi(n, \ell - 1) - 1 & + & \Phi(n-1, \ell - 1). \end{array}$$

We will consider three cases.

(1) When $n = 1$ and $\forall \ell$, then the condition reduces to $N^\Lambda(z^\ell) > 1$. That means that there exist at least two distinct loss vectors, which must differ at some sample point $z^\star \in \{z_1, \cdots, z_\ell\}$, i.,e. $L(z^\star; \alpha_1) = 0$ and $L(z^\star; \alpha_2) = 1$ for some α_1, α_2. Consequently, $N^\Lambda(z^\star) = 2$, and the result holds.

(2) When $\ell < n$, the condition is never valid since $N^(z^\ell) \leq 2^\ell$, whereas the condition states that:

$$N^\Lambda(z^\ell) > \sum_{i=0}^{n-1} \left(\begin{array}{c} \ell \\ i \end{array} \right) = \sum_{i=0}^{\ell-1} \left(\begin{array}{c} \ell \\ i \end{array} \right) = 2^\ell.$$

Therefore the result holds trivially.

(3) In all other interesting cases, we proceed by double induction.

Induction on n

The basis of the outer induction is case (1): the lemma is true for $n = 1$, $\forall \ell$. The induction is to assume the lemma is true for $n \leq N$, $\forall \ell$, and to prove that it is true for $n = N + 1$, $\forall \ell$.

Induction on ℓ

The basis of the inner induction is case (2), the lemma is true for $n = N + 1$ and $\ell < n$. The induction is to assume the lemma is true for $n \leq N + 1$, $\ell \leq L$, and to prove that it is true for $n = N + 1$, $\ell = L + 1$. The condition is that for some selection z^{L+1} the following is satisfied:

$$N^{\Lambda}(z^{L+1}) > \Phi(N + 1, L + 1).$$

Given this, we need to show that there exists a subset $\{z_1^\star, \cdots, z_{N+1}^\star\}$ such that:

$$N^{\Lambda}(z^{\star(N+1)}) = 2^{N+1}.$$

Consider the subset $\{z_1, \cdots, z_L\}$. There can be two cases:

(a) $N^{\Lambda}(z_1, \cdots, z_L) > \Phi(N + 1, L)$.
By prior induction, this case is already established to validate the lemma. Thus, there exists a subset of $\{z_1, \cdots, z_L\}$ of length $N + 1$, such that $N^{\Lambda}(z^{\star(N+1)}) = 2^{N+1}$, and the lemma holds.

(b) $N^{\Lambda}(z_1, \cdots, z_L) \leq \Phi(N + 1, L)$.
Categorize all subsets $A \subset \{z_1, \cdots, z_L\}$ into two types: type I (decoupled from z_L) and type II (coupled with z_L):

I. If there exists α such that $L(A, \alpha) = 1$ (read $L(z, \alpha) = 1\ \forall z \in A$), $L(A^c, \alpha) = 1$ and $L(z_L, \alpha) = 1$ and there exists α' such that $L(A, \alpha') = 1$, $L(A^c, \alpha') = 1$ and $L(z_L, \alpha') = 0$ (i.e. selection of A is decoupled from selection of z_L.)

II. If either there exists α such that $L(A, \alpha) = 1$, $L(A^c, \alpha) = 1$ and $L(z_L, \alpha) = 1$ or there exists α' such that $L(A, \alpha') = 1$, $L(A^c, \alpha') = 1$ and $L(z_L, \alpha') = 0$ but not both (i.e. selection of A is coupled with selection of z_L.)

Let K_1=# of subsets of type I, and K_2=# of subsets of type II. The following equalities follow:

$$\begin{aligned} N^{\Lambda}(z_1, \cdots, z_L) &= K_1 + K_2, \\ N^{\Lambda}(z_1, \cdots, z_L, z_{L+1}) &= 2K_1 + K_2, \text{ thus :} \\ N^{\Lambda}(z_1, \cdots, z_{L+1}) &= N^{\Lambda}(z_1, \cdots, z_L) + K_1. \end{aligned}$$

Now define $A_\alpha = \{z | z \in \{z_1, \cdots, z_L\} \text{ and } L(z; \alpha) = 1\}$, and let us restrict the parameter space to only those that generate subsets of type I:

$$\tilde{\Lambda} = \{\alpha \mid A_\alpha \text{ is of type I}\}$$

Clearly, it follows that $K_1 = N^{\tilde{\Lambda}}(z_1, \cdots, z_L)$.

Note that, up to the induced steps, the validity of the lemma is confirmed for any parameter space, including $\tilde{\Lambda}$. Therefore we have that:

- If $K_1 = N^{\tilde{\Lambda}}(z_1, \cdots, z_L) > \Phi(N, L)$, then by induction there exists a subset $\{z_1^\star, \cdots, z_N^\star\}$ such that $N^{\tilde{\Lambda}}(z_1^\star, \cdots, z_N^\star) = 2^N$. It follows that $N^{\Lambda}(z_1^\star, \cdots, z_N^\star) = 2^N$, since $2^N \geq N^{\Lambda}(z^{\star N}) \geq N^{\tilde{\Lambda}}(z^{\star N}) = 2^N$.

  This means that A_α, $\alpha \in \tilde{\Lambda}$ can select any subset of $z^{\star N}$. Also, these A_α are decoupled from z_{L+1}, i.e. type I (for z^L and z_{L+1}), so it follows that $A'_\alpha = A_\alpha \cap \{z_1^\star, \cdots, z_N^\star\}$ are decoupled from z_{L+1}, i.e. type I (for $z^{\star N}$ and z_{L+1}). These two observations imply that all subsets of $\{z_1^\star, \cdots, z_N^\star\}$ are decoupled from z_{L+1}, they are all type I, and thus K'_1=# of subsets of type I (for $z^{\star N}$ and z_{L+1}) $= N^{\Lambda}(z^{\star N}) = 2^N$.

  Therefore the third equality above applies, and the lemma is true, since:

$$N^{\Lambda}(z_1^\star, \cdots, z_N^\star, z_{L+1}) = N^{\Lambda}(z_1^\star, \cdots, z_N^\star) + K'_1 = 2^{N+1}.$$

- If $K_1 = N^{\tilde{\Lambda}}(z_1, \cdots, z_L) \leq \Phi(N, L)$, then:

$$\begin{array}{rcll} N^{\Lambda}(z_1, \cdots, z_{L+1}) & = & N^{\Lambda}(z_1, \cdots, z_L) + K_1 & \text{(third equality)} \\ & \leq & N^{\Lambda}(z_1, \cdots, z_L) + \Phi(N, L) & \text{(current hypothesis)} \\ & \leq & \Phi(N+1, L) + \Phi(N, L) & \text{(case (b))} \\ & = & \Phi(N+1, L+1) & \text{(property of } \Phi\text{)} \end{array}$$

This contradicts the hypothesized condition, and this case is thus impossible.

This completes the induction on ℓ.

It follows that the induction on n is also complete.

□

Lemma 2. *If for some n:*

$$\sup_{z_1, \cdots, z_{n+1}} N^{\Lambda}(z_1, \cdots, z_{n+1}) \neq 2^{n+1},$$

then for all $\ell > n$

$$\sup_{z_1, \cdots, z_{\ell}} N^{\Lambda}(z_1, \cdots, z_{\ell}) \leq \Phi(n+1, \ell).$$

Proof. Assume the converse, for the sake of contradiction, i.e. there exists $z_1, \cdots, z_\ell$ such that:

$$N^{\Lambda}(z_1, \cdots, z_{\ell}) > \Phi(n+1, \ell)$$

Then, by Lemma 1, there exists a subset $z_1^\star, \cdots, z_{n+1}^\star$ such that $N^{\Lambda}(z_1^\star, \cdots, z_{n+1}^\star) = 2^{n+1}$, which contradicts our hypothesis, as desired. □

Corollary. *If h=VC dimension of Λ, then*

$$\begin{array}{rcl} \sup_{z_1, \cdots, z_{\ell}} N^{\Lambda}(z_1, \cdots, z_{\ell}) & \leq & \Phi(h+1, \ell) \\ & = & \displaystyle\sum_{i=0}^{h} \left(\begin{array}{c} l \\ i \end{array} \right), \end{array}$$

which is nothing but the number of subsets of a set of size ℓ, of size at most h.

Lemma 3. *We have:*

$$\Phi(n, \ell) \overset{(1)}{\leq} 1.5 \frac{\ell^{n-1}}{(n-1)!} \overset{(2)}{<} 1.5 \left[\frac{e\ell}{n-1} \right]^{n-1}.$$

Proof. The following is a concise sketch:

(1) By induction.

(2) Using Stirling's approximation.

□

Finally, Lemmas 2 and 3 combined establish the theorem.