

# Machine Learning Theory (CS 6783)

Lecture 7 : Growth Function, Massart's Finite Lemma, VC Dimension

## 1 Recap

1. We have the following bound for the ERM  $\hat{y}$

$$\mathbb{E}_S \left[ L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} L_D(f) - L_S(f) \right]$$

2. We also showed that by symmetrization

$$\mathbb{E}_S \left[ L_D(\hat{y}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq \frac{2}{n} \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right]$$

3. For binary classification problem since we can write  $\ell(y', y) = \mathbf{1}\{y' \neq y\} = \frac{1-y'y}{2}$ , we have

$$\frac{2}{n} \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right] = \frac{1}{n} \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right]$$

Why do we care? What has this bought us?

## 2 Effective size of function class on Data

Why is the introduction of Rademacher averages important ? To analyze the term,  $\mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right]$  consider the inner expectation, that is conditioned on sample consider the term  $\mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right]$ . Note that  $\frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)$  is still average of 0 mean random variables (conditioned on data) and we can apply Hoeffding bound for each fixed  $f \in \mathcal{F}$  individually. Now  $\mathcal{F}$  might be an infinite class, but, conditioned on input instances  $(x_1, y_1), \dots, (x_n, y_n)$ , one can ask, what is the size of the projection set

$$\mathcal{F}_{|x_1, \dots, x_n} = \{f(x_1), \dots, f(x_n) : f \in \mathcal{F}\}$$

For any binary class  $\mathcal{F}$ , first note that this set can have a maximum cardinality of  $2^n$  however it could be much smaller. In fact we can have,

$$\mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right] = \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{\mathbf{f} \in \mathcal{F}_{|x_1, \dots, x_n}} \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(\mathbf{f}[t], y_t) \right] \leq \mathbb{E}_S \left[ \sqrt{\frac{\log |\mathcal{F}_{|x_1, \dots, x_n}|}{n}} \right]$$

where the last step is using the Lemma 3 which we shall prove in a bit. Now one can define the growth function for a hypothesis class  $\mathcal{F}$  as follows.

$$\Pi_{\mathcal{F}}(\mathcal{F}, n) = \sup \{ |\mathcal{F}_{|x_1, \dots, x_n}| : x_1, \dots, x_n \in \mathcal{X} \}$$

**Example : thresholds**

What does the growth function of the class of threshold function look like ?

Well sort any given  $n$  points in ascending order, using thresholds, we can get at most  $n + 1$  possible labeling on the  $n$  points. Hence  $\Pi_{\mathcal{F}}(n) = n + 1$ . From this we conclude that for the learning thresholds problem,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{\log(n)}{n}}$$

### 3 Massart's Finite Lemma

**Lemma 1.** For any set  $V \subset \mathbb{R}^n$  :

$$\frac{1}{n} \mathbb{E}_{\epsilon} \left[ \sup_{\mathbf{v} \in V} \sum_{t=1}^n \epsilon_t \mathbf{v}[t] \right] \leq \frac{1}{n} \sqrt{2 \left( \sup_{\mathbf{v} \in V} \sum_{t=1}^n \mathbf{v}^2[t] \right) \log |V|}$$

*Proof.*

$$\begin{aligned} \sup_{\mathbf{v} \in V} \sum_{t=1}^n \epsilon_t \mathbf{v}[t] &= \frac{1}{\lambda} \log \left( \sup_{\mathbf{v} \in V} \exp \left( \lambda \sum_{t=1}^n \epsilon_t \mathbf{v}[t] \right) \right) \\ &\leq \frac{1}{\lambda} \log \left( \sum_{\mathbf{v} \in V} \exp \left( \lambda \sum_{t=1}^n \epsilon_t \mathbf{v}[t] \right) \right) \\ &= \log \left( \sum_{\mathbf{v} \in V} \prod_{t=1}^n \exp(\lambda \epsilon_t \mathbf{v}[t]) \right) \end{aligned}$$

Taking expectation w.r.t. Rademacher random variables,

$$\begin{aligned} \mathbb{E}_{\epsilon} \left[ \sup_{\mathbf{v} \in V} \sum_{t=1}^n \epsilon_t \mathbf{v}[t] \right] &\leq \frac{1}{\lambda} \mathbb{E}_{\epsilon} \left[ \log \left( \sum_{\mathbf{v} \in V} \prod_{t=1}^n \exp(\lambda \epsilon_t \mathbf{v}[t]) \right) \right] \\ &\leq \frac{1}{\lambda} \log \left( \sum_{\mathbf{v} \in V} \prod_{t=1}^n \mathbb{E}_{\epsilon_t} [\exp(\lambda \epsilon_t \mathbf{v}[t])] \right) \\ &= \frac{1}{\lambda} \log \left( \sum_{\mathbf{v} \in V} \prod_{t=1}^n \frac{e^{\lambda \mathbf{v}[t]} + e^{-\lambda \mathbf{v}[t]}}{2} \right) \end{aligned}$$

For any  $x$ ,  $\frac{e^x + e^{-x}}{2} \leq e^{x^2/2}$

$$\begin{aligned} &\leq \frac{1}{\lambda} \log \left( \sum_{\mathbf{v} \in V} e^{\lambda^2 \sum_{t=1}^n \mathbf{v}^2[t]/2} \right) \\ &\leq \frac{1}{\lambda} \log \left( |V| e^{\lambda^2 \sup_{\mathbf{v} \in V} (\sum_{t=1}^n \mathbf{v}^2[t])/2} \right) \\ &= \frac{\log |V|}{\lambda} + \frac{\lambda \sup_{\mathbf{v} \in V} (\sum_{t=1}^n \mathbf{v}^2[t])}{2} \end{aligned}$$

Choosing  $\lambda = \sqrt{\frac{2 \log |V|}{\sup_{\mathbf{v} \in V} (\sum_{t=1}^n \mathbf{v}^2[t])}}$  completes the proof. □

## 4 Growth Function and VC dimension

Growth function is defined as,

$$\Pi(\mathcal{F}, n) = \max_{x_1, \dots, x_n} |\mathcal{F}|_{x_1, \dots, x_n}|$$

Clearly we have from the previous results on bounding minimax rates for statistical learning in terms of cardinality of growth function that :

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{2 \log \Pi(\mathcal{F}, n)}{n}}$$

Note that  $\Pi(\mathcal{F}, n)$  is at most  $2^n$  but it could be much smaller. In general how do we get a handle on growth function for a hypothesis class  $\mathcal{F}$ ? Is there a generic characterization of growth function of a hypothesis class ?

**Definition 1.** *VC dimension of a binary function class  $\mathcal{F}$  is the largest number of points  $d = \text{VC}(\mathcal{F})$ , such that*

$$\Pi_{\mathcal{F}}(d) = 2^d$$

*If no such  $d$  exists then  $\text{VC}(\mathcal{F}) = \infty$*

If for any set  $\{x_1, \dots, x_n\}$  we have that  $|\mathcal{F}|_{x_1, \dots, x_n}| = 2^n$  then we say that such a set is shattered. Alternatively VC dimension is the size of the largest set that can be shattered by  $\mathcal{F}$ . We also define VC dimension of a class  $\mathcal{F}$  restricted to instances  $x_1, \dots, x_n$  as

$$\text{VC}(\mathcal{F}; x_1, \dots, x_n) = \max \left\{ t : \exists i_1, \dots, i_t \in [n] \text{ s.t. } \left| \mathcal{F}|_{x_{i_1}, \dots, x_{i_t}} \right| = 2^t \right\}$$

That is the size of the largest shattered subset of  $n$ . Note that for any  $n \geq \text{VC}(\mathcal{F})$ ,  $\sup_{x_1, \dots, x_n} \text{VC}(\mathcal{F}|_{x_1, \dots, x_n}) = \text{VC}(\mathcal{F})$ .

**Eg. Thresholds** One point can be shattered, but two points cannot be shattered. Hence VC dimension is 1. (If we allow both threshold to right and left, VC dimension is 2).

**Eg. Spheres Centered at Origin in  $d$  dimensions** one point can be shattered. But even two can't be shattered. VC dimension is 1!

**Eg. Half-spaces** Consider the hypothesis class where all points to the left (or right) of a hyperplane in  $\mathbb{R}^d$  are marked positive and the rest negative. In 1 dimension this is threshold both to left and right. VC dimension is 2. In  $d$  dimensions, think of why  $d + 1$  points can be shattered.  $d + 2$  points can't be shattered. Hence VC dimension is  $d + 1$ .

**Claim 2.** *Learnability with binary hypothesis class  $\mathcal{F}$  implies  $\text{VC}(\mathcal{F}) < \infty$ .*

*Proof.* First note that learnability in the statistical learning framework implies learnability in the realizable PAC setting. Hence to prove the claim, it suffices to show that if a hypothesis class has infinite VC dimension, then it is not even learnable in the realizable PAC setting. To this end, assume that a hypothesis class  $\mathcal{F}$  has infinite VC dimension. This means that for any  $n$ , we can find

$2n$  points  $x_1, \dots, x_{2n}$  that are shattered by  $\mathcal{F}$ . Also drawn  $y_1, \dots, y_{2n} \in \{\pm 1\}$  Rademacher random variables. Let  $D$  be the uniform distribution over the  $2n$  instance pairs  $(x_1, y_1), \dots, (x_{2n}, y_{2n})$ . Notice that since  $x_1, \dots, x_{2n}$  are shattered by  $\mathcal{F}$ , we are indeed in the realizable PAC setting for any choice of  $y$ 's. Now assume we get  $n$  input instances drawn iid from this distribution. Clearly in this sample of size  $n$ , we can at most witness  $n$  unique instances. Let us denote  $J \subset [2n]$  as the indices of the  $2n$  instances witnessed in the draw of  $n$  samples  $S$ . Clearly  $|J| \leq n$ . Hence we have,

$$\begin{aligned}
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) &\geq \sup_{x_1, \dots, x_{2n}} \inf_{\hat{y}} \mathbb{E}_{y_1, \dots, y_{2n}} \mathbb{E}_S \left[ \frac{1}{2n} \sum_{j=1}^{2n} \mathbf{1}_{\{\hat{y}(x_i) \neq y_i\}} \right] \\
&= \frac{1}{2n} \sup_{x_1, \dots, x_{2n}} \inf_{\hat{y}} \mathbb{E}_{y_1, \dots, y_{2n}} \mathbb{E}_J \left[ \sum_{i \in J} \mathbf{1}_{\{\hat{y}(x_i) \neq y_i\}} + \sum_{i \in [2n] \setminus J} \mathbf{1}_{\{\hat{y}(x_i) \neq y_i\}} \right] \\
&\geq \frac{1}{2n} \sup_{x_1, \dots, x_{2n}} \inf_{\hat{y}} \min_{J \subset [2n]: |J| \leq n} \mathbb{E}_{y_1, \dots, y_{2n}} \left[ \sum_{i \in [2n] \setminus J} \mathbf{1}_{\{\hat{y}(x_i) \neq y_i\}} \right] \\
&= \frac{1}{4n} \min_{J \subset [2n]: |J| \leq n} |[2n] \setminus J| = \frac{n}{4n} = \frac{1}{4}
\end{aligned}$$

□

**Lemma 3** (VC'71/Sauer'72/Shelah'72). *For any class  $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$  with  $\text{VC}(\mathcal{F}) = d$ , we have that,*

$$\Pi(\mathcal{F}, n) \leq \sum_{i=0}^d \binom{n}{i}$$

*Proof.* For notational ease let  $g(d, n) = \sum_{i=0}^d \binom{n}{i}$ . We want to prove that  $\Pi(\mathcal{F}, n) \leq g(d, n) = g(d, n-1) + g(d-1, n-1)$ . We prove this one by induction on  $n+d$ .

**Base case :** We need to consider two base cases. First, note that when VC dimension  $d = 0$ , then clearly for any  $x, x' \in \mathcal{X}$ ,  $f(x) = f(x')$  and so we can conclude that for such a class  $\mathcal{F}$  effectively contains only one function and so  $\Pi(\mathcal{F}, n) = g(0, n) = 1$ . On the other hand, note that for any  $d \geq 1$ , if VC dimension of the function class  $\mathcal{F}$  is  $d$  then it can at least shatter 1 point and so  $\Pi(\mathcal{F}, 1) = g(d, 1) = 2$ . These form our base case.

**Induction :** Assume that the statement holds for any class  $\mathcal{F}$  with VC dimension  $d' \leq d$  and any  $n' \leq n-1$  that  $\Pi(\mathcal{F}, n') \leq g(d', n')$ . We shall prove that in this case, for any  $\mathcal{F}$  with VC dimension  $d' \leq d$ ,  $\Pi(\mathcal{F}, n) \leq g(d', n)$  and similarly for any  $n' \leq n$ , and for any  $\mathcal{F}$  with VC dimension at most  $d+1$ ,  $\Pi(\mathcal{F}, n') \leq g(d+1, n')$ .

To this end, consider any class  $\mathcal{F}$  of VC dimension at most  $d'$  and consider any set of  $n$  instances  $x_1, \dots, x_n$ . Define hypothesis class

$$\tilde{\mathcal{F}} = \{f \in \mathcal{F} : \exists f' \in \mathcal{F} \text{ s.t. } f(x_n) \neq f'(x_n), \forall i < n, f(x_i) = f'(x_i)\}$$

That is the hypothesis class consisting of all functions that have a pair with same exact value of  $x_1, \dots, x_{n-1}$  but opposite sign only on  $x_n$ . We first claim that,

$$|\mathcal{F}_{|x_1, \dots, x_n}| = |\mathcal{F}_{|x_1, \dots, x_{n-1}}| + |\tilde{\mathcal{F}}_{|x_1, \dots, x_{n-1}}|$$

This is because  $\tilde{\mathcal{F}}_{|x_1, \dots, x_{n-1}}$  are exactly the elements that need to be counted twice (once for + and once for -). We also claim that  $\text{VC}(\tilde{\mathcal{F}}; x_1, \dots, x_{n-1}) \leq d' - 1$  because if not, by definition of  $\tilde{\mathcal{F}}$  we know that  $\tilde{\mathcal{F}}$  can shatter  $x_n$  and so we will have that

$$\text{VC}(\tilde{\mathcal{F}}; x_1, \dots, x_n) = \text{VC}(\tilde{\mathcal{F}}; x_1, \dots, x_{n-1}) + 1 = d' + 1$$

This is a contradiction as  $\tilde{\mathcal{F}}$  is a subset of  $\mathcal{F}$  which itself has only VC dimension at most  $d'$ . Thus we conclude that for any class  $\mathcal{F}$  of VC dimension at most  $d'$ ,

$$\begin{aligned} \Pi(\mathcal{F}, n) &= \sup_{x_1, \dots, x_n} |\mathcal{F}_{|x_1, \dots, x_n}| \\ &\leq \sup_{x_1, \dots, x_n} \left\{ |\mathcal{F}_{|x_1, \dots, x_{n-1}}| + |\tilde{\mathcal{F}}_{|x_1, \dots, x_{n-1}}| \right\} \end{aligned}$$

where  $\text{VC}(\tilde{\mathcal{F}}; x_1, \dots, x_{n-1})$  is at most  $d - 1$ . Using the above bound, the inductive hypothesis and the fact that  $g(d', n) = g(d', n - 1) + g(d' - 1, n - 1)$ , we conclude that for any class  $\mathcal{F}$  with VC dimension at most  $d' \leq d$ ,

$$\begin{aligned} \Pi(\mathcal{F}, n) &\leq \sup_{x_1, \dots, x_n} \left\{ |\mathcal{F}_{|x_1, \dots, x_{n-1}}| + |\tilde{\mathcal{F}}_{|x_1, \dots, x_{n-1}}| \right\} \\ &\leq g(d', n - 1) + g(d' - 1, n - 1) = g(d', n) \end{aligned}$$

Similarly for any  $n' \leq n$ , and for any  $\mathcal{F}$  with VC dimension at most  $d + 1$ , we can show by repeatedly using the inductive hypothesis, starting from  $n' = 2$  up until  $n' = n$  that for any  $\Pi(\mathcal{F}, n') \leq g(d + 1, n')$ . This concludes out induction.  $\square$

**Remark 4.1.** Note that  $\sum_{i=0}^d \binom{n}{i} \leq \left(\frac{n}{d}\right)^d$ . Hence we can conclude that for any binary classification problem with hypothesis class  $\mathcal{F}$  in the statistical learning setting, if  $\text{VC}_{\mathcal{F}} \leq d$  then,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \frac{1}{n} \sup_D \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] \leq \sqrt{\frac{d \log \left(\frac{n}{d}\right)}{n}}$$

The above statement basically implies that if a binary hypothesis class  $\mathcal{F}$  has finite VC dimension, then it is learnable in the statistical learning (agnostic PAC) framework.