Machine Learning Theory (CS 6783)

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

1 Recap

1. Bound below holds for Empirical Risk Minimizer (ERM) as well.

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

2. Binary classification problem

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

2 Growth Function

Why is the introduction of Rademacher averages important? To analyze the term, consider the inner expectation, that is conditioned on sample. Consider the term \( \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \). Note that \( \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(x_t) \) is still average of 0 mean random variables 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 \( x_1, \ldots, x_n \), one can ask, what is the size of the projection set

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

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,

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

where the last step is using the Lemma 1 which we shall prove in a bit. Now one can define the growth function for a hypothesis class \( \mathcal{F} \) as follows.

\[ \Pi_{\mathcal{F}}(\mathcal{F}, n) = \sup \{|\mathcal{F}|_{x_1, \ldots, x_n} : x_1, \ldots, x_n \in \mathcal{X} \} \]

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,

\[ V_{\text{stat}}(\mathcal{F}) \leq \sqrt{\log(n)} \]
3 Massart’s Finite Lemma

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

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

Proof.

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

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

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

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

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

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

4 Growth Function and VC dimension

Growth function is defined as,

$$
\Pi(\mathcal{F}, n) = \max_{x_1, \ldots, x_n} |\mathcal{F}|_{x_1, \ldots, x_n}
$$
Clearly we have from the previous results on bounding minimax rates for statistical learning in terms of cardinality of growth function that:

\[ \sqrt{n} \text{stat}(F) \leq \sqrt{2 \log \Pi(F, n)} \]

Note that \( \Pi(F, n) \) 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 \( F \)? Is there a generic characterization of growth function of a hypothesis class?

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

\[ \Pi(F, d) = 2^d \]

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

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

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

That is the size of the largest shattered subset of \( n \). Note that for any \( n \geq \text{VC}(F) \),

\[ \sup_{x_1, \ldots, x_n} \text{VC}(F_{x_1, \ldots, x_n}) = \text{VC}(F). \]

**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 \( d + 1 \).

**Lemma 2** (VC’71/Sauer’72/Shelah’72). For any class \( F \subset \{\pm 1\}^X \) with \( \text{VC}(F) = d \), we have that,

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

**Proof.** For notational ease let \( g(d, n) = \sum_{i=0}^{d} \binom{n}{i} \). We want to prove that \( \Pi(F, n) \leq g(d, n) = g(d, n-1) + g(d-1, n-1) \). 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 \( x, x' \in X \), \( f(x) = f(x') \) and so we can conclude that for such a class \( F \) effectively contains only one function and so \( \Pi(F, n) = g(0, n) = 1 \). 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 the 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 \( x_1, \ldots, x_n \). 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 \( x_1, \ldots, x_{n-1} \) but opposite sign only on \( x_n \). We first claim that,

\[
|\mathcal{F}|_{x_1, \ldots, x_n} = |\mathcal{F}|_{x_1, \ldots, x_{n-1}} + |\tilde{\mathcal{F}}|_{x_1, \ldots, x_{n-1}}
\]

This is because \( \tilde{\mathcal{F}} \) are exactly the elements that need to be counted twice (once for + and once for −). We also claim that VC(\( \tilde{\mathcal{F}} ; x_1, \ldots, x_{n-1} \)) \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

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

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 \( d' \),

\[
\Pi(\mathcal{F}, n) = \sup_{x_1, \ldots, x_n} |\mathcal{F}|_{x_1, \ldots, x_n}
\]

\[
\leq \sup_{x_1, \ldots, x_n} \{ |\mathcal{F}|_{x_1, \ldots, x_{n-1}} + |\tilde{\mathcal{F}}|_{x_1, \ldots, x_{n-1}} \}
\]

where VC(\( \tilde{\mathcal{F}} ; x_1, \ldots, x_{n-1} \)) 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 \( d' \leq d \),

\[
\Pi(\mathcal{F}, n) \leq \sup_{x_1, \ldots, x_n} \{ |\mathcal{F}|_{x_1, \ldots, x_{n-1}} + |\tilde{\mathcal{F}}|_{x_1, \ldots, x_{n-1}} \}
\]

\[
\leq g(d', n - 1) + g(d' - 1, n - 1) = g(d', n)
\]

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 \( \sum_{i=0}^{d} \binom{n}{i} \leq \left( \frac{n}{2} \right)^d \). Hence we can conclude that for any binary classification problem with hypothesis class \( \mathcal{F} \) in the statistical learning setting, if VC\( \mathcal{F} \) \leq d \) then,

\[
\gamma_n^{\text{stat}}(\mathcal{F}) \leq \frac{1}{n} \sup_{D} \mathbb{E}_D \mathbb{E}_x \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i) \right] \leq \sqrt{\frac{d \log \left( \frac{n}{4} \right)}{n}}
\]

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.