1 Recap

1. For the ERM we have,

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

RHS above is the Rademacher complexity of the loss composed with function class \( F \)

2. This is useful because conditioned on data, we can get bounds that depend on effective size of \( F \) on data \( x_1, \ldots, x_n \).

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

where \( F_{|x_1, \ldots, x_n} = \{(f(x_1), \ldots, f(x_n)) : f \in F\} \)

3. Eg. threshold is learnable and effective size on \( n \) points is at most \( n + 1 \) but \( F \) is uncountably infinite.

4. Massart’s finite lemma implies:

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

2 Massart’s Finite Lemma

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

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

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

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

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

Since log is a concave function, by Jensen’s inequality, Expected log is upper bounded by log of expectation and so:

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

For any \( x \), \( e^x + e^{-x} \leq e^{x^2/2} \)

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

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

\[\square\]

3 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:

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

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

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

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

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

1. To show \( \text{VC}(\mathcal{F}) \geq d \) show that you can at least pick \( d \) points \( x_1, \ldots, x_d \) that can be shattered.
2. To show that \( \text{VC}(\mathcal{F}) \leq d \) show that no configuration of \( d + 1 \) points can be shattered.

**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. VC dimension is \( d + 1 \).

**Lemma 2** (VC’71 (originially 64!)/Sauer’72/Shelah’72). For any class \( \mathcal{F} \subset \{\pm 1\}^X \) with \( \text{VC}(\mathcal{F}) = d \), we have that,

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

**Remark 3.1.** 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 \( \mathcal{F} \),

\[ V_n^{\text{stat}}(\mathcal{F}) \leq \frac{1}{n} \sup_D \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \leq \sqrt{\frac{\text{VC}(\mathcal{F}) \log \left( \frac{n}{\text{VC}(\mathcal{F})} \right)}{n}} \]
Hence, if a binary hypothesis class $\mathcal{F}$ has finite VC dimension, then it is learnable in the statistical learning (agnostic PAC) framework. $\log(n/\text{VC}(\mathcal{F}))$ in the above bound can be removed.

**Proof of VC Lemma.** 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 $x_1, \ldots, x_n$. 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,

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

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

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

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

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

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 our induction. □