Machine Learning Theory (CS 6783)

Lecture 20 : Sequential Rademacher Complexity and Properties

1 Recap

- Using minimax theorem repeatedly and the idea of conditional symmetrization we showed:

$$\begin{aligned}\mathcal{V}_n^{sq}(\mathcal{F}) &= \frac{1}{n}\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(Y)} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \inf_{\hat{y}_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t\sim p_t} [\ell(\hat{y}_t, y_t)] - \ell(f(x_t), y_t)\right] \\ &\le \frac{1}{n}\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(Y)} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{y_t\sim p_t} [\ell(f(x_t), y_t)] - \ell(f(x_t), y_t)\right] \\ &\le \frac{2}{n}\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{y_t\in Y} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right]\end{aligned}$$

- Sequential Rademacher Complexity upper bounds the minimax rate for online learning

$$\mathcal{V}_n^{sq}(\mathcal{F}) = \frac{2}{n} \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}_t(\epsilon)), \mathbf{y}_t(\epsilon))\right] = 2\mathcal{R}_n^{sq}(\ell\circ\mathcal{F})$$

- For convex Lipschitz loss and Binary losses $\mathcal{R}_n^{sq}(\ell\circ\mathcal{F}) \le L\ \mathcal{R}_n^{sq}(\mathcal{F})$.
- Properties of classical Rademacher complexity and hold for sequential version.
- For absolute loss (also for hinge and zero one), $\mathcal{V}_n^{sq}(\mathcal{F}) \ge \mathcal{R}_n^{sq}(\mathcal{F})$

1.1 Finite Lemma

Lemma 1. *For any set V of real valued trees of depth n,*

$$\frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{\mathbf{v}\in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)\right] \le \frac{1}{n}\sqrt{2\left(\sup_{\mathbf{v}\in V} \max_{\epsilon\in\{\pm1\}^n} \sum_{t=1}^{n} \mathbf{v}_t^2(\epsilon)\right)\log|V|}$$

Proof idea. Similar to the iid version of finite lemma except on trees. We start with replacing max with soft-max and using Jensen.

$$\mathbb{E}_\epsilon \left[\sup_{\mathbf{v}\in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)\right] \le \inf_{\lambda>0} \frac{1}{\lambda} \log\left(\sum_{\mathbf{v}\in V} \mathbb{E}_\epsilon \left[\exp\left(\lambda \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)\right)\right]\right)$$

For $t \in \{0, \ldots, n-1\}$, define $A^t : \{\pm 1\}^t \to \mathbb{R}$ by $A^t(\epsilon_1, \ldots, \epsilon_t) = \max_{\epsilon_{t+1}, \ldots, \epsilon_n} \exp\left\{\frac{\lambda^2}{2} \sum_{s=t+1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\}$ and $A^n(\epsilon_1, \ldots, \epsilon_n) = 1$. We have that for any $t \in \{1, \ldots, n\}$

$$
\begin{aligned}
&\mathbb{E}_{\epsilon_t}\left[\exp\left(\lambda \sum_{s=1}^{t} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^t(\epsilon_1, \ldots, \epsilon_t)\right] \\
&= \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, +1) + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, -1)\right) \\
&\leq \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \max_{\epsilon_t \in \{\pm 1\}} A^t(\epsilon_1, \ldots, \epsilon_t) \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})}\right) \\
&\leq \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^{t-1}(\epsilon_1, \ldots, \epsilon_{t-1})
\end{aligned}
$$

where in the last step we used the inequality $(e^a + e^{-a})/2 \leq e^{a^2/2}$. Thus we can conclude that

$$
\mathbb{E}_\epsilon\left[\sup_{\mathbf{v} \in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)\right] \leq \inf_{\lambda > 0}\left\{\frac{\log |V|}{\lambda} + \frac{1}{\lambda} \log\left(\max_{\mathbf{v} \in V} \max_{\epsilon} \exp\left\{\frac{\lambda^2}{2} \sum_{s=1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\}\right)\right\}
$$

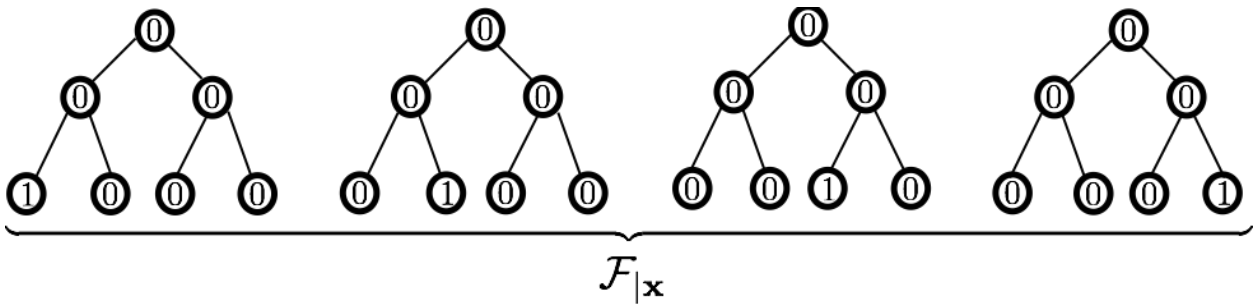
□

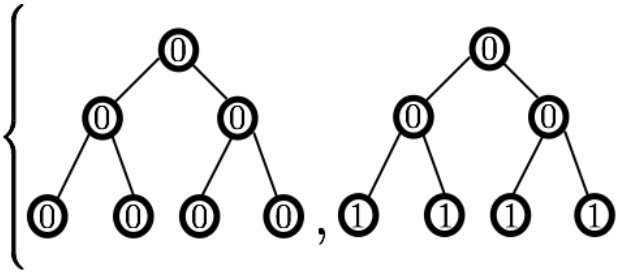
2 Growth Function and Covering Number

In the iid case we looked at (effective) cardinality $|\mathcal{F}_{|x_1, \ldots, x_n}|$. For online learning should we look at $\mathcal{F}_{|\mathbf{x}}$? ($\mathcal{F}_{|\mathbf{x}}$ is the set of real valued trees got by projecting $\mathcal{F}$ on to tree $\mathbf{x}$, that is $\mathcal{F}_{|\mathbf{x}} = f(\mathbf{x}) : f \in \mathcal{F}$). Is this the right quantity ? Clearly,

$$
\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon))\right] = \mathbb{E}_\epsilon\left[\sup_{\mathbf{v} \in \mathcal{F}_{|\mathbf{x}}} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)\right]
$$

But is the size of $\mathcal{F}_{|\mathbf{x}}$ the right quantity?



$$V = \left\{ \right.$$

$$\left. \right\}$$

$$\mathbb{E}_\epsilon \left[\sup_{\mathbf{v} \in \mathcal{F}_{|\mathbf{x}}} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon) \right] = \mathbb{E}_\epsilon \left[\sup_{\mathbf{v} \in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon) \right]$$

Definition 1. *Given an $\mathcal{X}$-valued tree $\mathbf{x}$, we say that the set V of real valued trees is an ℓ_p cover of $\mathcal{F}$ on $\mathbf{x}$ at scale α if*

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \ s.t. \ \left(\frac{1}{n} \sum_{t=1}^{n} |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)|^p \right)^{1/p} \leq \alpha$$

Also define covering numbers

$$\mathcal{N}_p^{sq}(\mathcal{F}, \alpha, \mathbf{x}) = \min\{|V| : V \text{ is a } \ell_p \text{ cover of } \mathcal{F} \text{ on } \mathbf{x} \text{ at scale } \alpha\}$$

and $\mathcal{N}_p^{sq}(\mathcal{F}, \alpha, n) = \sup_{\mathbf{x}} \mathcal{N}_p(\mathcal{F}, \alpha, \mathbf{x})$

In other words, can we replace $\mathcal{F}_{|\mathbf{x}}$ by a set V such that, for any tree in $\mathcal{F}_{|\mathbf{x}}$ and any path, there exists a tree in V that is close on the same path. From this definition of covering number, for binary class $\mathcal{F}$, we can now define growth function as $\Pi^{sq}(\mathcal{F}, n) = \mathcal{N}_p^{sq}(\mathcal{F}, 0, n)$.

2.1 Sequential Pollard Bound

Lemma 2. *We have the following bound on the sequential Rademacher complexity :*

$$\mathcal{R}_n^{sq}(\mathcal{F}) \leq \inf_{\beta \geq 0} \left\{ \beta + \sqrt{\frac{2 \log \mathcal{N}_1^{sq}(\mathcal{F}, \beta, n)}{n}} \right\}$$

The proof is very similar to the iid version. Only we use the sequential version of the finite lemma.

2.2 Sequential Dudley Integral Bound

Lemma 3. *We have the following bound on the sequential Rademacher complexity :*

$$\mathcal{R}_n^{sq}(\mathcal{F}) \leq \inf_{\beta \geq 0} \left\{ 4\beta + 12 \int_{\beta}^{1/2} \sqrt{\frac{\log \mathcal{N}_2^{sq}(\mathcal{F}, \delta, n)}{n}} d\delta \right\} =: \mathcal{D}_n^{sq}(\mathcal{F})$$

While this proof has the same chaining idea at its heart, does require a bit of work.

3 Littlestone Dimension and Sequential Fat-shattering Dimension

Definition 2. *Given a binary function class $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ and a $\mathcal{X}$-valued tree $\mathbf{x}$, we say that $\mathcal{F}$ shatters $\mathbf{x}$ if*

$$\forall \epsilon \in \{\pm 1\}^n, \exists f_\epsilon \in \mathcal{F} \text{ s.t. } f_\epsilon(\mathbf{x}_t(\epsilon)) = \epsilon_t$$

Further we define the Littlestone dimension of a class $\mathcal{F}$ as :

$$ldim(\mathcal{F}) = \sup\{d : \exists \mathcal{X}\text{-valued tree } \mathbf{x} \text{ of depth } d \text{ that is shattered by } \mathcal{F}\}$$

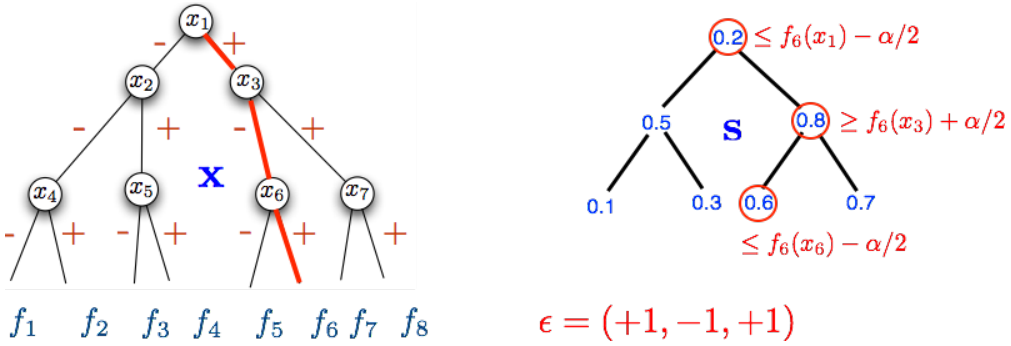
Let us also define the scale sensitive version of this combinatorial dimension, the sequential fat-shattering dimension before we proceed to use these quantities.

Definition 3. *Given $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and a $\mathcal{X}$-valued tree $\mathbf{x}$, we say that $\mathcal{F}$ shatters $\mathbf{x}$ at scale α, if there exists a $\mathbb{R}$-valued witness tree $\mathbf{s}$ such that,*

$$\forall \epsilon \in \{\pm 1\}^n, \exists f_\epsilon \in \mathcal{F} \text{ s.t. } (f(\mathbf{x}_t(\epsilon) - \mathbf{s}_t(\epsilon))\epsilon_t \geq \alpha/2$$

Further we define the sequential fat-shattering dimension of a class $\mathcal{F}$ as :

$$\text{fat}_\alpha^{sq}(\mathcal{F}) = \sup\{d : \exists \mathcal{X}\text{-valued tree } \mathbf{x} \text{ of depth } d \text{ that is } \alpha - shatteredby \mathcal{F}\}$$



Examples : Thresholds have infinite Littlestone dimension. Monotonic functions have infinite fat-shattering dimension at scale 1/2 for example.

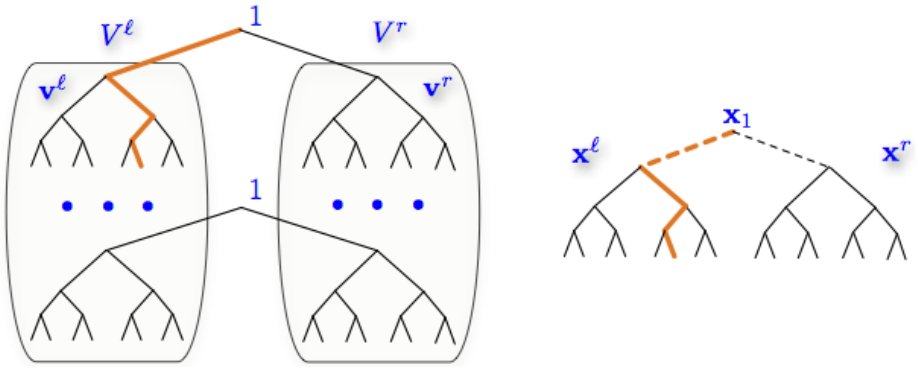
Lemma 4 (Sequential VC-Sauer-Shelah Lemma)**.** *For any $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and any $\gamma \in (0,1)$*

$$\mathcal{N}_\infty^{sq}(\mathcal{F}, \gamma, n) \leq \left(\frac{n}{\gamma}\right)^{\text{fat}_\gamma^{sq}(\mathcal{F})}$$

Proof is induction based like in the iid setting, however the way we induct is different. I only outline the proof here. Take $\mathcal{F}$ to be binary for simplicity. We show that

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

Base case is obvious. To show the inductive hypothesis, given a tree we split $\mathcal{G}$ into two classes based on value at root ($+1$ or -1). Now one of these subclasses has to have Little stone dimension strictly smaller than $ldim$ of base class. If not, using this root as root of a new tree of depth $ldim + 1$, we can get a tree of depth $ldim + 1$ that is shattered. (the tree version proof in a sense is easier and more apparent!)



Remark 3.1 (Relating the Various Complexities)**.** *We already have,*

$$\mathcal{R}_n^{sq}(\mathcal{F}) \leq \mathcal{D}_n^{sq}(\mathcal{F}) \leq \inf_{\beta \geq 0} \left\{ 4\beta + 12 \int_\beta^{1/2} \sqrt{\frac{\mathrm{fat}_\delta^{sq}(\mathcal{F}) \log(n/\delta)}{n}} d\delta \right\} =: \mathcal{C}_n^{sq}(\mathcal{F})$$

But we can also show that for any $\beta > 0$:

$$\min\{n, \mathrm{fat}_\beta^{sq}(\mathcal{F})\} \leq \frac{32 n \mathcal{R}_n^{sq}(\mathcal{F})^2}{\beta^2}$$

From this we can conclude that

$$\mathcal{R}_n^{sq}(\mathcal{F}) \geq \Omega^* \left(\mathcal{C}_n^{sq}(\mathcal{F}) \right)$$

4 Putting It All Together

Theorem 5. *For any real valued hypothesis class $\mathcal{F}$, and supervised statistical learning problem with absolute loss (also for squared loss, logistic loss,...), the following are equivalent :*

1. *$\mathcal{F}$ is online learnable ($\mathcal{V}_n^{sq}(\mathcal{F}) \to 0$)*
2. *$\mathcal{R}_n^{sq}(\mathcal{F}) \to 0$*
3. *$\mathcal{D}_n^{sq}(\mathcal{F}) \to 0$*
4. *$\forall \gamma > 0$, $\mathrm{fat}_\gamma^{sq} < \infty$*

Further all three sequential complexity measures can be used to get bounds that are at most log factors of each other and to the minimax rate (absolute loss). These complexity measures also provide necessary an sufficient conditions and near optimal rates of convergence for uniform law of large numbers for martingales. That is for following convergence

$$\sup_{\mathbf{P} \in \Delta(\mathcal{X})^n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{t=1}^n \left(f(X_t) - \mathbb{E}\left[f(X_t) | X_1, \ldots, X_{t-1} \right] \right) \right| \right] \to 0$$