

Machine Learning Theory (CS 6783)

Lecture 15 : Sequential Rademacher Complexity : Properties

1 Recap

•

$$\begin{aligned} \mathcal{V}_n^{sq}(\mathcal{F}) &\leq \frac{1}{n} \left\| \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(Y)} \mathbb{E}_{y_t \sim p_t} \right\|_{t=1}^n \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \mathbb{E}_{y_t \sim p_t} [\ell(f(x_t), y_t)] - \ell(f(x_t), y_t) \right] \\ &\leq \frac{2}{n} \left\| \sup_{x_t \in \mathcal{X}} \sup_{y_t \in Y} \mathbb{E}_{\epsilon_t} \right\|_{t=1}^n \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right] \\ &= \frac{2}{n} \sup_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \ell(f(\mathbf{x}_t(\epsilon)), \mathbf{y}_t(\epsilon)) \right] \end{aligned}$$

- Define sequential Rademacher complexity of a class $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$ as

$$\mathcal{R}_n^{sq}(\mathcal{G}) = \frac{1}{n} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon} \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t g(\mathbf{z}_t(\epsilon)) \right]$$

We have, $\mathcal{V}_n^{sq}(\mathcal{F}) \leq 2\mathcal{R}_n^{sq}(\ell \circ \mathcal{F})$

2 Lower Bound on Online Learning

Let $\mathcal{Y} = [-1, 1]$ and $\ell(y', y) = |y' - y|$.

Claim 1.

$$\mathcal{V}_n^{sq}(\mathcal{F}) \geq \mathcal{R}_n^{sq}(\mathcal{F})$$

Proof. We start with the equality of the minimax rate from two lectures ago. And for the lower bound we specifically choose the distributions on y 's to be fair coin flip with $\{\pm 1\}$ outcomes. Hence,

$$\begin{aligned} \mathcal{V}_n^{sq}(\mathcal{F}) &= \frac{1}{n} \left\| \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(Y)} \mathbb{E}_{y_t \sim p_t} \right\|_{t=1}^n \left[\sum_{t=1}^n \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{y_t \sim p_t} [|\hat{y}_t - y_t|] - \inf_{f \in \mathcal{F}} \sum_{t=1}^n |f(x_t) - y_t| \right] \\ &\geq \frac{1}{n} \left\| \sup_{x_t \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \right\|_{t=1}^n \left[\sum_{t=1}^n \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} [|\hat{y}_t - \epsilon_t|] - \inf_{f \in \mathcal{F}} \sum_{t=1}^n |f(x_t) - \epsilon_t| \right] \\ &\geq \frac{1}{n} \left\| \sup_{x_t \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \right\|_{t=1}^n \left[n - \inf_{f \in \mathcal{F}} \sum_{t=1}^n (1 - f(x_t)\epsilon_t) \right] \\ &= \frac{1}{n} \left\| \sup_{x_t \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \right\|_{t=1}^n \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] = \mathcal{R}_n^{sq}(\mathcal{F}) \end{aligned}$$

□

3 Properties of Sequential Rademacher Complexity

Proposition 2. For any classes $\mathcal{G}$, $\mathcal{H}$ mapping instances in $\mathcal{Z}$ to reals :

1. If $\mathcal{H} \subset \mathcal{G}$, then $\mathcal{R}_n^{sq}(\mathcal{H}) \leq \mathcal{R}_n^{sq}(\mathcal{G})$
2. For any fixed function $h : \mathcal{Z} \mapsto \mathbb{R}$, $\mathcal{R}_n^{sq}(\mathcal{G} + h) = \mathcal{R}_n^{sq}(\mathcal{G})$
3. $\mathcal{R}_n^{sq}(\text{cvx}(\mathcal{G})) = \mathcal{R}_n^{sq}(\mathcal{G})$
4. $\mathcal{R}_n^{sq}(\mathcal{H})(\mathcal{G} + \mathcal{H}) = \mathcal{R}_n^{sq}(\mathcal{G}) + \mathcal{R}_n^{sq}(\mathcal{H})$

Proof for the above properties are identical to proofs for the classical Rademacher complexity version from Lecture 7.

Below we prove a proposition that turns out to be helpful for removing the loss function from the complexity measure in many cases.

Proposition 3. Let $\mathbf{s}$ be any $\{-1, 1\}$ valued tree of depth n , then,

$$\frac{1}{n} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon} \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t \mathbf{s}_t(\epsilon) g(\mathbf{z}_t(\epsilon)) \right] = \frac{1}{n} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon} \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t g(\mathbf{z}_t(\epsilon)) \right]$$

Proof. The statement follows from a very simple observation. Consider any $a \in \{\pm 1\}$ and any arbitrary function $\Phi : \pm 1 \mapsto \mathbb{R}$. We have that

$$\mathbb{E}_{\epsilon \sim \text{Unif}\{\pm 1\}} [\Phi(\epsilon \cdot a)] = \frac{\Phi(a) + \Phi(-a)}{2} = \frac{\Phi(1) + \Phi(-1)}{2} = \mathbb{E}_{\epsilon \sim \text{Unif}\{\pm 1\}} [\Phi(\epsilon)]$$

We can use the above to conclude the proposition. Let $\mathbf{s}$ be any $\{\pm 1\}$ -valued tree and $\mathbf{z}$ any $\mathcal{Z}$ -valued tree. For each t , Given $\epsilon_1, \dots, \epsilon_{t-1}$, define

$$\Phi_t(a) = \left\langle \sup_{z_j \in \mathcal{Z}} \mathbb{E}_{\epsilon'_j} \right\rangle_{j=t+1}^n \left[\sup_{g \in \mathcal{G}} \left\{ \sum_{i=1}^{t-1} \epsilon_i \mathbf{s}_i(\epsilon) g(\mathbf{z}_i(\epsilon)) + a \cdot g(\mathbf{z}_t(\epsilon)) + \sum_{i=t+1}^n \epsilon'_i g(z_i) \right\} \right]$$

Note that given any $\mathbf{s}$ and $\mathbf{z}$,

$$\mathbb{E}_{\epsilon} [\Phi_n(\mathbf{s}_n(\epsilon) \cdot \epsilon_n)] = \frac{1}{n} \mathbb{E}_{\epsilon} \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t \mathbf{s}_t(\epsilon) g(\mathbf{z}_t(\epsilon)) \right]$$

Also note that $\Phi_0 = \left\langle \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\epsilon'_t} \right\rangle_{t=1}^n [\sup_{g \in \mathcal{G}} \{\sum_{t=1}^n \epsilon'_t g(z_t)\}] = \mathcal{R}_n^{sq}(\mathcal{G})$ also note that,

$$\begin{aligned} \mathbb{E}_{\epsilon_t} [\Phi_t(\epsilon_t)] &= \mathbb{E}_{\epsilon_t} \left[\left\langle \sup_{z_j \in \mathcal{Z}} \mathbb{E}_{\epsilon'_j} \right\rangle_{j=t+1}^n \left[\sup_{g \in \mathcal{G}} \left\{ \sum_{i=1}^{t-1} \epsilon_i \mathbf{s}_i(\epsilon) g(\mathbf{z}_i(\epsilon)) + \epsilon_t \cdot g(\mathbf{z}_t(\epsilon)) + \sum_{i=t+1}^n \epsilon'_i g(z_i) \right\} \right] \right] \\ &\leq \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\epsilon_t} \left[\left\langle \sup_{z_j \in \mathcal{Z}} \mathbb{E}_{\epsilon'_j} \right\rangle_{j=t+1}^n \left[\sup_{g \in \mathcal{G}} \left\{ \sum_{i=1}^{t-1} \epsilon_i \mathbf{s}_i(\epsilon) g(\mathbf{z}_i(\epsilon)) + \epsilon_t \cdot g(z_t) + \sum_{i=t+1}^n \epsilon'_i g(z_i) \right\} \right] \right] = \Phi_{t-1}(\mathbf{s}_{t-1}(\epsilon) \cdot \epsilon_{t-1}) \end{aligned}$$

Now since we already showed that for any $a \in \{\pm 1\}$, $\Phi_t(a \cdot \epsilon_t) = \mathbb{E}_{\epsilon_t} [\Phi_t(\epsilon_t)]$, we have that ,

$$\begin{aligned} \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t \mathbf{s}_t(\epsilon) g(\mathbf{z}_t(\epsilon)) \right] &= \mathbb{E}_\epsilon [\Phi_n(\mathbf{s}_n(\epsilon) \cdot \epsilon_n)] = \mathbb{E}_\epsilon [\Phi_n(\epsilon_n)] \leq \mathbb{E}_\epsilon [\Phi_{n-1}(\mathbf{s}_{n-1}(\epsilon) \cdot \epsilon_{n-1})] \\ &= \dots \leq \Phi_0 = \frac{1}{n} \sup_{\mathbf{z}} \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}} \sum_{t=1}^n \epsilon_t g(\mathbf{z}_t(\epsilon)) \right] \end{aligned}$$

□

- Binary classification : $\ell(y', y) = \mathbf{1}_{\{y' \neq y\}} = \frac{1-yy'}{2}$ hence $\mathbf{R}_n = \frac{1}{2n} (\sum_{t=1}^n \hat{y}_t y_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^n f(x_t) y_t)$

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

- Convex Lipchitz loss : $\mathcal{Y} \subset \mathbb{R}$, $\ell(\hat{y}, y)$ is convex and L -Lipschitz in $\hat{y}$. First note that since loss is convex, no randomization required.

$$\mathbf{R}_n = \frac{1}{n} \left(\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) \right) \leq \frac{1}{n} \left(\sum_{t=1}^n \partial \ell(\hat{y}_t, y_t) \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \partial \ell(\hat{y}_t, y_t) f(x_t) \right)$$

Since y_t is picked after adversary sees $\hat{y}_t$, think of adversary, instead of picking y_t picks $\partial_t = \partial \ell(\hat{y}_t, y_t) \in [-L, L]$. Thus the value of the original learning problem is bounded by minimax rate of the learning problem with linear loss $\partial_t \cdot \hat{y}_t$. Hence,

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

where in the above ∂ is a $[-L, L]$ -valued tree. Since term is convex in ∂ it is maximized at vertex $\{-L, L\}$ valued tree. Now using above proposition we can get rid of the gradient tree.

3.1 Finite Lemma

Lemma 4. For any set V of real valued trees of depth n ,

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

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

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

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

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

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

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

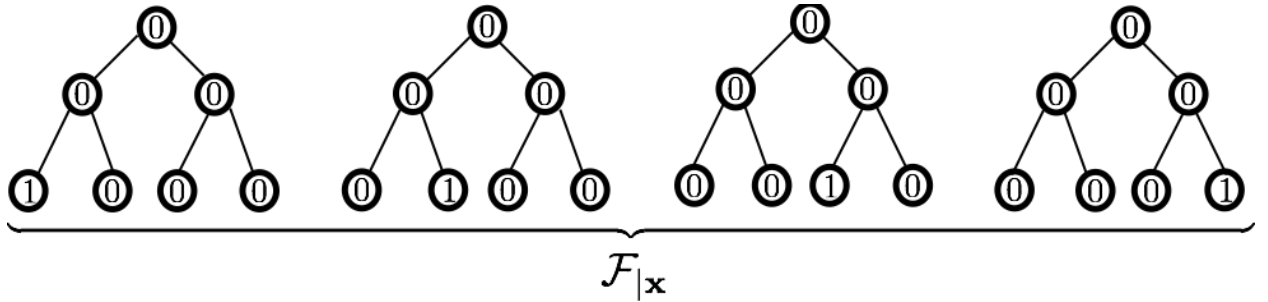
□

4 Growth Function and Covering Number

In the iid case we looked at (effective) cardinality $|\mathcal{F}_{|x_1, \dots, x_n}|$. 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?



$$V = \left\{ \begin{array}{c} \text{0} \\ \swarrow \quad \searrow \\ \text{0} \quad \text{0} \\ \swarrow \quad \searrow \quad \swarrow \quad \searrow \\ \text{0} \quad \text{0} \quad \text{0} \quad \text{0} \end{array} , \begin{array}{c} \text{0} \\ \swarrow \quad \searrow \\ \text{0} \quad \text{0} \\ \swarrow \quad \searrow \quad \swarrow \quad \searrow \\ \text{1} \quad \text{1} \quad \text{1} \quad \text{1} \end{array} \right\}$$

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