# 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:

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

- Further we also showed

$$V_n((x_1,y_1),\ldots,(x_t,y_t)) = \left\langle\!\!\left\langle \sup_{x_j\in\mathcal{X}} \sup_{p_j\in\Delta(Y)} \mathop{\mathbb{E}}_{y_j\sim p_j} \right\rangle\!\!\right\rangle_{j=t+1}^{n} \left[\sum_{j=t+1}^{n} \inf_{\hat{y}_j\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_j\sim p_j}\left[\ell(\hat{y}_j, y_j)\right] - \inf_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i)\right]$$

# 2 Sequential Rademacher Complexity

The above complexity can be equivalently written as follows.

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

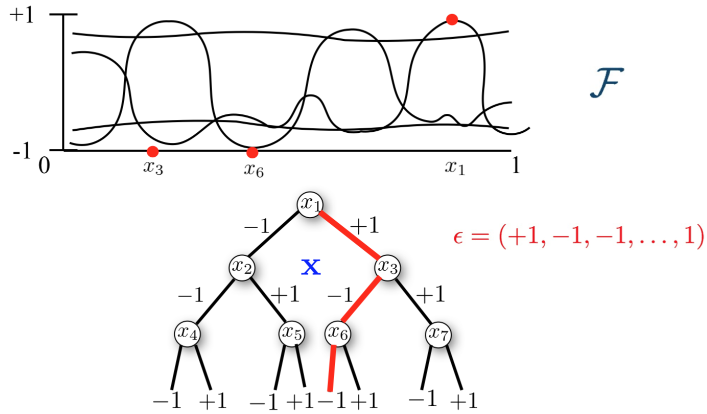
Where $\mathbf{x}$ and $\mathbf{y}$ are $\mathcal{X}$ and $\mathcal{Y}$ valued complete binary tree of depth $n$. That is, for instance $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where each $\mathbf{x}_t : \{\pm-1\}^{t-1} \mapsto \mathcal{X}$.

In general for a given function class $\mathcal{G}$ on space $\mathcal{Z}$ to reals we define below the sequential Rademacher complexity.

**Definition 1.** *Given a class $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, we define the sequential Rademacher complexity of the class $\mathcal{G}$ as,*

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

Pictorially, we can view the Rademacher complexity as :




To see that the two forms are equivalent, note that, given any trees $\mathbf{x}$ and $\mathbf{y}$, note that

$$
\begin{aligned}
&\sup_{\substack{x_1\in\mathcal{X}\\ y_1\in Y}} \mathbb{E}_{\epsilon_1} \dots \sup_{\substack{x_n\in\mathcal{X}\\ y_n\in Y}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right] \\
&\qquad \geq \sup_{\substack{x_1\in\mathcal{X}\\ y_1\in\mathcal{Y}}} \mathbb{E}_{\epsilon_1} \dots \sup_{\substack{x_{n-1}\in\mathcal{X}\\ y_{n-1}\in\mathcal{Y}}} \mathbb{E}_{\epsilon_{n-1}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n-1} \epsilon_t \ell(f(x_t), y_t) + \ell(f(\mathbf{x}_n(\epsilon), \mathbf{y}_n(\epsilon)))\right] \\
&\qquad \geq \sup_{\substack{x_1\in\mathcal{X}\\ y_1\in\mathcal{Y}}} \mathbb{E}_{\epsilon_1} \dots \sup_{\substack{x_t\in\mathcal{X}\\ y_t\in\mathcal{Y}}} \mathbb{E}_{\epsilon_{t+1:n}}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{t} \epsilon_i \ell(f(x_i), y_i) + \sum_{j=t+1}^{n} \ell(f(\mathbf{x}_j(\epsilon), \mathbf{y}_j(\epsilon)))\right] \\
&\qquad \geq \mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(\epsilon), \mathbf{y}_t(\epsilon)))\right]
\end{aligned}
$$

Since the above statement holds for any trees $\mathbf{x}$ and $\mathbf{y}$ we can take the supremum over the trees. On the other hand, define a pair of tree $\mathbf{x}^*$ and $\mathbf{y}^*$ as follows :

$$
\mathbf{x}_1^* = \operatorname*{argmax}_{x\in\mathcal{X}} \sup_{y_1\in\mathcal{Y}} \mathbb{E}_{\epsilon_1}\left[\left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ y_t\in Y}} \mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=2}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right]\right]
$$

(and similarly define $\mathbf{y}_1^*$) and subsequently, given each $\epsilon_{1:t-1}$ define

$$
\mathbf{x}_t^*(\epsilon_{1:t-1}) = \operatorname*{argmax}_{x\in\mathcal{X}} \sup_{y_t\in\mathcal{Y}} \mathbb{E}_{\epsilon_t}\left[\left\langle\!\!\left\langle \sup_{\substack{x_j\in\mathcal{X}\\ y_j\in Y}} \mathbb{E}_{\epsilon_j}\right\rangle\!\!\right\rangle_{j=t+1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{t-1} \epsilon_i \ell(f(\mathbf{x}_i(\epsilon)), \mathbf{y}_i(\epsilon)) + \sum_{j=t}^{n} \epsilon_j \ell(f(x_j), y_j)\right]\right]
$$

Clearly by definition of these trees,

$$
\sup_{\substack{x_1\in\mathcal{X}\\ y_1\in Y}} \mathbb{E}_{\epsilon_1} \dots \sup_{\substack{x_n\in\mathcal{X}\\ y_n\in Y}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right] \leq \mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t^*(\epsilon), \mathbf{y}_t^*(\epsilon)))\right]
$$

Since we have both inequalities we conclude that the two forms are equivalent.

## 3 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,

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

□

## 4 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. $\mathcal{R}_n^{sq}(\mathrm{cvx}(\mathcal{G})) = \mathcal{R}_n^{sq}(\mathcal{G})$
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

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

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 $\epsilon_1, \ldots, \epsilon_{t-1}$, define

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

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

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

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

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

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

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

□

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

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

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

## 4.1 Finite Lemma

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

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

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

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

For $t \in \{0, \ldots, n-1\}$, define $A^t : \{\pm1\}^t \to \mathbb{R}$ 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 $t \in \{1, \ldots, n\}$

$$\begin{aligned}
&\mathbb{E}_{\epsilon_t}\left[\exp\left(\lambda\sum_{s=1}^t \epsilon_s\mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^t(\epsilon_1, \ldots, \epsilon_t)\right] \\
&= \exp\left(\lambda\sum_{s=1}^{t-1} \epsilon_s\mathbf{v}_s(\epsilon_{1:s-1})\right) \times \left(\frac{1}{2}e^{\lambda\mathbf{v}_t(\epsilon_{1:t-1})}A^t(\epsilon_1, \ldots, \epsilon_{t-1}, +1) + \frac{1}{2}e^{-\lambda\mathbf{v}_t(\epsilon_{1:t-1})}A^t(\epsilon_1, \ldots, \epsilon_{t-1}, -1)\right) \\
&\leq \exp\left(\lambda\sum_{s=1}^{t-1} \epsilon_s\mathbf{v}_s(\epsilon_{1:s-1})\right) \times \max_{\epsilon_t\in\{\pm1\}} A^t(\epsilon_1, \ldots, \epsilon_t)\left(\frac{1}{2}e^{\lambda\mathbf{v}_t(\epsilon_{1:t-1})} + \frac{1}{2}e^{-\lambda\mathbf{v}_t(\epsilon_{1:t-1})}\right) \\
&\leq \exp\left(\lambda\sum_{s=1}^{t-1} \epsilon_s\mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^{t-1}(\epsilon_1, \ldots, \epsilon_{t-1})
\end{aligned}$$

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

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

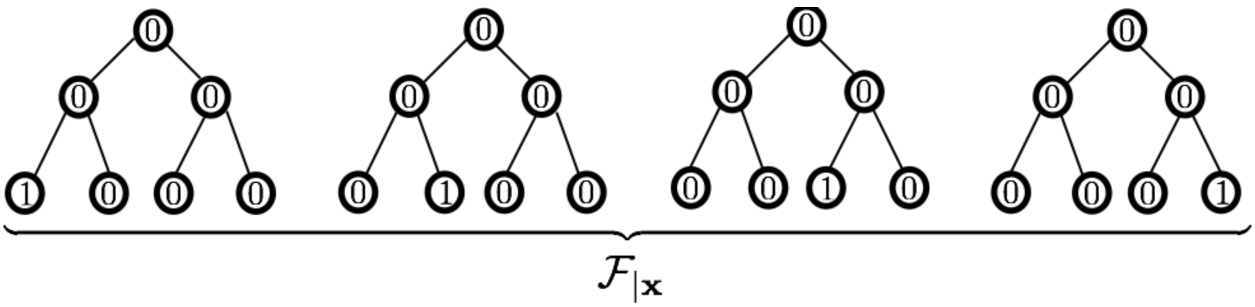
□

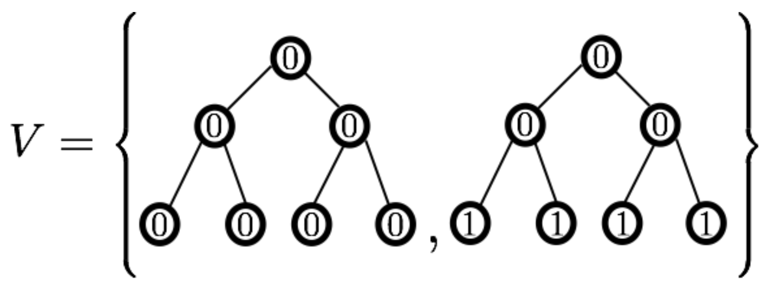
# 5 Growth Function and Covering Number

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?




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