1 Recap

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

\[
V_{sq}^n(F) = \frac{1}{n} \left( \sup_{x_t \in X} \sup_{p_t \in \Delta(Y)} \mathbb{E}_{y_t \sim p_t} \left[ \ell(\hat{y}_t, y_t) - \ell(f(x_t), y_t) \right] \right)^n \\
\leq \frac{1}{n} \left( \sup_{x_t \in X} \sup_{y_t \in Y} \mathbb{E}_{\epsilon_t \sim \Delta(Y)} \left[ \ell(f(x_t), y_t) - \ell(f(x_t), y_t) \right] \right)^n \\
\leq \frac{2}{n} \left( \sup_{x_t \in X} \sup_{y_t \in Y} \mathbb{E}_{\epsilon_t} \left[ \ell(f(x_t), y_t) \right] \right)^n \\
\leq \frac{2}{n} \left( \sup_{x_t \in X} \sup_{y_t \in Y} \mathbb{E}_{\epsilon_t} \left[ \ell(f(x_t), y_t) \right] \right)^n \\
\leq \frac{2}{n} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)
\]

- Further we also showed

\[
V_n((x_1, y_1), \ldots, (x_t, y_t)) = \left( \sup_{x_j \in X} \sup_{p_j \in \Delta(Y)} \mathbb{E}_{y_j \sim p_j} \left[ \ell(\hat{y}_j, y_j) - \inf_{f \in F} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right] \right)^n \\
\leq \frac{2}{n} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)
\]

2 Sequential Rademacher Complexity

The above complexity can be equivalently written as follows.

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

Where \( x \) and \( y \) are \( X \) and \( Y \) valued complete binary tree of depth \( n \). That is, for instance \( x = (x_1, \ldots, x_n) \) where each \( x_t : \{\pm -1\}^{t-1} \rightarrow X \).

In general for a given function class \( G \) on space \( Z \) to reals we define below the sequential Rademacher complexity.

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

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

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

\[
\sup_{\mathbf{x}_1 \in \mathcal{X}} \ldots \sup_{\mathbf{x}_n \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t) \right] \\
\geq \sup_{\mathbf{x}_1 \in \mathcal{X}} \ldots \sup_{\mathbf{x}_{t-1} \in \mathcal{X}} \mathbb{E}_{\epsilon_{t-1}} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n-1} \epsilon_t \ell(f(x_t), y_t) + \ell(f(x_n(\epsilon), y_n(\epsilon))) \right] \\
\geq \sup_{\mathbf{x}_1 \in \mathcal{X}} \ldots \sup_{\mathbf{x}_{t-1} \in \mathcal{X}} \mathbb{E}_{\epsilon_{t-1}} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{t} \epsilon_i \ell(f(x_i), y_i) + \sum_{j=t+1}^{n} \ell(f(x_j(\epsilon), y_j(\epsilon))) \right] \\
\geq \mathbb{E}_{\epsilon} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t(\epsilon), y_t(\epsilon))) \right]
\]

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:

\[
\mathbf{x}_1^* = \arg\max_{\mathbf{x} \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \mathbb{E}_{\epsilon_1} \left[ \left\langle \sup_{\mathbf{x}_2 \in \mathcal{X}} \mathbb{E}_{\epsilon_2} \right\rangle \ldots \sup_{\mathbf{x}_t \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{t} \epsilon_i \ell(f(x_i), y_i) \right] \right]
\]

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

\[
\mathbf{x}_t^*(\epsilon_{1:t-1}) = \arg\max_{\mathbf{x} \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \left[ \left\langle \sup_{\mathbf{x}_t \in \mathcal{X}} \mathbb{E}_{\epsilon_t} \right\rangle \sup_{\mathbf{x}_j \in \mathcal{X}} \mathbb{E}_{\epsilon_j} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{j} \epsilon_i \ell(f(x_i), y_i) + \sum_{j+t}^{n} \epsilon_j \ell(f(x_j), y_j) \right] \right]
\]

Clearly by definition of these trees,

\[
\sup_{\mathbf{x}_1 \in \mathcal{X}} \ldots \sup_{\mathbf{x}_n \in \mathcal{X}} \mathbb{E}_{\epsilon_n} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t) \right] \leq \mathbb{E}_{\epsilon} \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t^*(\epsilon), y_t^*(\epsilon))) \right]
\]

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.

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

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,

$$\mathcal{V}^\text{sq}_n(F) = \frac{1}{n} \sup_{x_t \in X} \mathbb{E}_{\epsilon_t} \left[ \sum_{t=1}^{n} \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \left[ |\hat{y}_t - y_t| - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} |f(x_t) - y_t| \right] \right]$$

$$\geq \frac{1}{n} \sup_{x_t \in X} \mathbb{E}_{\epsilon_t} \left[ \sum_{t=1}^{n} \inf_{\hat{y}_t \in \mathcal{Y}} \mathbb{E}_{\epsilon_t} \left[ |\hat{y}_t - \epsilon_t| - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} |f(x_t) - \epsilon_t| \right] \right]$$

$$= \frac{1}{n} \sup_{x_t \in X} \mathbb{E}_{\epsilon_t} \left[ \sum_{t=1}^{n} \epsilon_t f(x_t) \right] = R^\text{sq}_n(F)$$

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 $R^\text{sq}_n(\mathcal{H}) \leq R^\text{sq}_n(\mathcal{G})$

2. For any fixed function $h : \mathcal{Z} \mapsto \mathbb{R}$, $R^\text{sq}_n(\mathcal{G} + h) = R^\text{sq}_n(\mathcal{G})$

3. $R^\text{sq}_n(\text{cvx}(\mathcal{G})) = R^\text{sq}_n(\mathcal{G})$

4. $R^\text{sq}_n(\mathcal{H})(\mathcal{G} + \mathcal{H}) = R^\text{sq}_n(\mathcal{G}) + R^\text{sq}_n(\mathcal{H})$

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 $s$ be any $\{\pm 1\}$ valued tree of depth $n$, then,

$$\frac{1}{n} \sup_{z} \mathbb{E}_{s} \left[ \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \epsilon_t s_t(\epsilon) g(z_t(\epsilon)) \right] = \frac{1}{n} \sup_{z} \mathbb{E}_{s} \left[ \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \epsilon_t g(z_t(\epsilon)) \right]$$
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 \( s \) be any \( \{\pm 1\} \)-valued tree and \( z \) any \( \mathbb{Z} \)-valued tree. For each \( t \), Given \( \epsilon_1, \ldots, \epsilon_{t-1} \), define

\[
\Phi_t(a) = \left( \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \right) \left( \sup_{i=1}^{t-1} \left[ \sum_{i=1}^{t-1} \epsilon_i s_i(\epsilon) (z_i(\epsilon)) + a \cdot g(z_i(\epsilon)) \right] \right)
\]

Note that given any \( s \) and \( z \),

\[
\mathbb{E}_{\epsilon} \left[ \Phi_n(s_n(\epsilon) \cdot \epsilon_n) \right] = \frac{1}{n} \mathbb{E}_{\epsilon} \left[ \sup_{g \in \mathcal{G}} \left\{ \sum_{t=1}^{n} \epsilon_t s_t(\epsilon) (z_t(\epsilon)) \right\} \right]
\]

Also note that \( \Phi_0 = \langle \left\{ \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \right\} \rangle = \mathcal{R}_n^{sq}(\mathcal{G}) \) also note that,

\[
\mathbb{E}_{\epsilon_t} [\Phi_t(\epsilon_t)] = \mathbb{E}_{\epsilon_t} \left[ \left( \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \right) \left( \sup_{i=1}^{t-1} \left[ \sum_{i=1}^{t-1} \epsilon_i s_i(\epsilon) (z_i(\epsilon)) + \epsilon_t \cdot g(z_i(\epsilon)) \right] \right) \right] \leq \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \left[ \left( \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \right) \left( \sup_{i=1}^{t-1} \left[ \sum_{i=1}^{t-1} \epsilon_i s_i(\epsilon) (z_i(\epsilon)) + \epsilon_t \cdot g(z_i(\epsilon)) \right] \right) \right] = \Phi_{t-1}(s_{t-1}(\epsilon) \cdot \epsilon_{t-1})
\]

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,

\[
\frac{1}{n} \mathbb{E}_{\epsilon} \left[ \sup_{g \in \mathcal{G}} \left\{ \sum_{t=1}^{n} \epsilon_t s_t(\epsilon) (z_t(\epsilon)) \right\} \right] = \mathbb{E}_{\epsilon} [\Phi_n(s_n(\epsilon) \cdot \epsilon_n)] = \mathbb{E}_{\epsilon} [\Phi_n(\epsilon_n)] \leq \mathbb{E}_{\epsilon} [\Phi_{n-1}(s_{n-1}(\epsilon) \cdot \epsilon_{n-1})] \leq \cdots \leq \Phi_0 = \frac{1}{n} \mathbb{E}_{\epsilon} \left[ \sup_{g \in \mathcal{G}} \left\{ \sum_{t=1}^{n} \epsilon_t s_t(\epsilon) (z_t(\epsilon)) \right\} \right]
\]

\[\Box\]

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

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

- 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.

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

4
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,

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

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 $n$,

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

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

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

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

$$\mathbb{E}_\epsilon \left[ \exp \left( \lambda \sum_{s=1}^{t} \epsilon_s v_s(\epsilon_{s-1}) \right) \times A^t(\epsilon_1, \ldots, \epsilon_t) \right]$$

$$= \exp \left( \lambda \sum_{s=1}^{t-1} \epsilon_s v_s(\epsilon_{s-1}) \right) \times \left( \frac{1}{2} e^{\lambda v_t(\epsilon_{t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, +1) + \frac{1}{2} e^{-\lambda v_t(\epsilon_{t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, -1) \right)$$

$$\leq \exp \left( \lambda \sum_{s=1}^{t-1} \epsilon_s v_s(\epsilon_{s-1}) \right) \times \max_{\epsilon_t \in \{\pm 1\}} A^t(\epsilon_1, \ldots, \epsilon_t) \left( \frac{1}{2} e^{\lambda v_t(\epsilon_{t-1})} + \frac{1}{2} e^{-\lambda v_t(\epsilon_{t-1})} \right)$$

$$\leq \exp \left( \lambda \sum_{s=1}^{t-1} \epsilon_s v_s(\epsilon_{s-1}) \right) \times A^{t-1}(\epsilon_1, \ldots, \epsilon_{t-1})$$

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

$$\mathbb{E}_\epsilon \left[ \sup_{v \in V} \sum_{t=1}^{n} \epsilon_t v_t(\epsilon) \right] \leq \inf_{\lambda > 0} \left\{ \frac{\log |V|}{\lambda} + \frac{1}{\lambda} \log \left( \max_{\epsilon \in \{\pm 1\}^n} \exp \left\{ \frac{\lambda^2}{2} \sum_{s=1}^{n} v_s(\epsilon_{s-1})^2 \right\} \right) \right\}$$
5 Growth Function and Covering Number

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

Is this the right quantity? Clearly,

$$\mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t(\epsilon)) \right] = \mathbb{E}_\epsilon \left[ \sup_{v \in \mathcal{F}_x} \sum_{t=1}^{n} \epsilon_t v_t(\epsilon) \right]$$

But is the size of $\mathcal{F}_x$ the right quantity?

$F_x$

$$V = \left\{ \begin{array}{c}
\begin{array}{c}
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{array}
, \\
\begin{array}{c}
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{array}
\end{array} \right\}$$

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