# Machine Learning Theory (CS 6783)

Lecture 13 : Online Learning, minimax value, sequential Rademacher complexity

## 1 Recap: Minimax Theorem

We shall use the celebrated minimax theorem as a key tool to bound the minimax rate for online learning problems. Below we state a generalization of Von Neuman's minimax theorem.

**Theorem 1** (Browein'14)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be Banach spaces. Let $A \subset \mathcal{A}$ be nonempty, weakly compact, and convex, and let $B \subset \mathcal{B}$ be nonempty and convex. Let $g : A \times B \mapsto \mathbb{R}$ be concave with respect to $b \in B$ and convex and lower-semicontinuous with respect to $a \in A$ and weakly continuous in $a$ when restricted to $A$. Then*

$$\sup_{b\in B} \inf_{a\in A} g(a,b) = \inf_{a\in A} \sup_{b\in B} g(a,b)$$

The above theorem states that under the right conditions, one can swap infimum and supremum. We shall use this in a sequential manner to swap the order of the learner and adversary and use this to get a handle on minimax rate for online learning. For instance using the above theorem, we can show that for any loss $\ell$, lower semicontinuous in its first argument, as long as $\mathcal{Y}$ is well behaved (compact for instance),

$$\inf_{q_t\in\Delta(\mathcal{Y})} \sup_{y_t\in\mathcal{Y}} \mathbb{E}_{\hat{y}_t\sim q_t} \left[\ell(\hat{y}_t, y_t) + \Phi(y_t)\right] = \sup_{p_t\in\Delta(\mathcal{Y})} \inf_{\hat{y}_t\in\mathcal{Y}} \mathbb{E}_{y_t\sim p_t} \left[\ell(\hat{y}_t, y_t) + \Phi(y_t)\right]$$

where $\Phi$ is some arbitrary function.

## 2 Minimax Rate for Online Learning

Recall that the minimax rate for an online learning problem can be written as :

$$\mathcal{V}_n^{sq} = \sup_{x_1\in\mathcal{X}} \inf_{q_1\in\Delta(\mathcal{Y})} \sup_{y_1\in\mathcal{Y}} \mathbb{E}_{\hat{y}_1\sim q_1} \cdots \sup_{x_n\in\mathcal{X}} \inf_{q_n\in\Delta(\mathcal{F})} \sup_{y_n\in\mathcal{Y}} \mathbb{E}_{\hat{y}_n\sim q_n} \left[\frac{1}{n}\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n \ell(f(x_t), y_t)\right]$$

That is in a sequential fashion, on each round, adversary picks the worst input instance $x_t \in \mathcal{X}$, The learner then picks the optimal $q_t \in \Delta(\mathcal{Y})$ the adversary then picks the worst outcome $y_t \in \mathcal{Y}$, then learner draws prediction $\hat{y}_t \sim q_t$ with the aim of learner to minimize regret and goal of adversary to maximize regret. We now introduce a shorthand notation. We shall use the notation $\langle\!\langle \mathbf{Operator}_t \rangle\!\rangle_{t=1}^n [\ldots]$ to refer to $\mathbf{Operator}_1 \mathbf{Operator}_2 \ldots \mathbf{Operator}_n [\ldots]$. Hence for instance,

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

We can also write the conditional value as

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

**Claim 2.**

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

*Proof.*

$$\begin{aligned}
n\mathcal{V}_n^{sq} &= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=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] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[ \sum_{t=1}^{n-1} \ell(\hat{y}_t, y_t) + \sup_{x_n \in \mathcal{X}} \inf_{q_n \in \Delta(\mathcal{Y})} \sup_{y_n \in \mathcal{Y}} \underbrace{\left\{ \mathop{\mathbb{E}}_{\hat{y}_n \sim q_n} [\ell(\hat{y}_n, y_n)] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\}}_{g(q_n, y_n)} \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[ \sum_{t=1}^{n-1} \ell(\hat{y}_t, y_t) + \sup_{x_n \in \mathcal{X}} \sup_{p_n \in \Delta(\mathcal{Y})} \inf_{\hat{y}_n \in \mathcal{Y}} \mathbb{E}_{y_n \sim p_n} \left[ \ell(\hat{y}_n, y_n) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[ \sum_{t=1}^{n-1} \ell(\hat{y}_t, y_t) + \sup_{x_n \in \mathcal{X}} \sup_{p_n \in \Delta(\mathcal{Y})} \inf_{\hat{y}_n \in \mathcal{Y}} \mathop{\mathbb{E}}_{y_n \sim p_n} [\ell(\hat{y}_n, y_n)] - \mathop{\mathbb{E}}_{y_n \sim p_n} \left[ \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[ \sum_{t=1}^{n-1} \ell(\hat{y}_t, y_t) + \sup_{x_n \in \mathcal{X}} \sup_{p_n \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_n \sim p_n} \left[ \inf_{\hat{y}_n \in \mathcal{Y}} \mathop{\mathbb{E}}_{y_n \sim p_n} [\ell(\hat{y}_n, y_n)] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \right] \\
&= \ldots \\
&= \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} [\ell(\hat{y}_t, y_t)] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \qquad = \left\langle\!\!\left\langle
\end{aligned}$$

Thus we have the claim. □

Notice that in the above claim, we have a distributions (possibly dependent) over instances but have essentially eliminated the role of the learner and moved to a completely stochastic object. From the above claim it is easy to show that the the minimax rate if governed by a quantity measuring rate of uniform convergence of class $\mathcal{F}$ over martingale difference sequences.

**Claim 3.**

$$\mathcal{V}_n^{sq} \leq \sup_{\mathbf{P} \in \Delta(\mathcal{X} \times \mathcal{Y})^n} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{t-1} [\ell(f(x_t), y_t)] - \ell(f(x_t), y_t) \right]$$

*where* $\mathbf{P}$ *is a joint distribution over the sequence of instances and* $\mathbb{E}_{t-1}[\cdot]$ *refers to the conditional expectation over instance* $(x_t, y_t)$ *given past instances* $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$

*Proof.*

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

□

# 3 Sequential Rademacher Complexity

In the statistical learning framework a key tool was symmetrization and the use of Rademacher complexity. With the use of Rademacher complexity we were able to move our focus on how the function class behaves on the entire space of instances to only how rich the class is effectively on samples of size $n$. The question now, is whether there is an analogue of this for uniform convergence over martingales. Surprisingly it turns out that there is and this complexity we shall refer to as sequential Rademacher complexity.

**Claim 4.**

$$\mathcal{V}_n^{sq} \le 2 \sup_{\substack{x_1\in\mathcal{X}\\ y_1\in Y}} \mathbb{E}_{\epsilon_1} \dots \sup_{\substack{x_n\in\mathcal{X}\\ y_n\in Y}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right]$$

*Proof.*

$$\begin{aligned}
n\mathcal{V}_n^{sq} &\le \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ 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] \\
&= \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ p_t\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sup_{\substack{x_n\in\mathcal{X}\\ p_n\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_n\sim p_n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{y'_t\sim p_t} [\ell(f(x_t), y'_t)] - \ell(f(x_t), y_t) + \mathop{\mathbb{E}}_{y'_n\sim p_n} [\ell(f(x_n), y'_n)] - \ell(f(x_n), y_n)\right]\right] \\
&\le \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ p_t\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sup_{\substack{x_n\in\mathcal{X}\\ p_n\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_n, y'_n\sim p_n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{y'_t\sim p_t} [\ell(f(x_t), y'_t)] - \ell(f(x_t), y_t) + \ell(f(x_n), y'_n) - \ell(f(x_n), y_n)\right]\right] \\
&= \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ p_t\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sup_{\substack{x_n\in\mathcal{X}\\ p_n\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_n, y'_n\sim p_n} \mathbb{E}_{\epsilon_n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{y'_t\sim p_t} [\ell(f(x_t), y'_t)] - \ell(f(x_t), y_t) + \epsilon_n(\ell(f(x_n), y'_n) - \ell(f(x_n), y_n))\right]\right] \\
&\le \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ p_t\in\Delta(Y)}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sup_{\substack{x_n\in\mathcal{X}\\ y_n, y'_n\in Y}} \mathbb{E}_{\epsilon_n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{y'_t\sim p_t} [\ell(f(x_t), y'_t)] - \ell(f(x_t), y_t) + \epsilon_n(\ell(f(x_n), y'_n) - \ell(f(x_n), y_n))\right]\right] \\
&\le \dots
\end{aligned}$$

proceeding in similar fashion

$$
\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ y_t,y_t'\in Y}} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t(\ell(f(x_t),y_t') - \ell(f(x_t),y_t))\right] \\
&\leq \left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ y_t,y_t'\in Y}} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t\ell(f(x_t),y_t') + \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} -\epsilon_t\ell(f(x_t),y_t))\right] \\
&\leq 2\left\langle\!\!\left\langle \sup_{\substack{x_t\in\mathcal{X}\\ y_t\in Y}} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t\ell(f(x_t),y_t)\right] \\
&= 2\sup_{\substack{x_1\in\mathcal{X}\\ y_1\in Y}} \mathbb{E}_{\epsilon_1} \ldots \sup_{\substack{x_n\in\mathcal{X}\\ y_n\in Y}} \mathbb{E}_{\epsilon_n} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t\ell(f(x_t),y_t)\right]
\end{aligned}
$$

□

The above complexity can be equivalently written as follows.

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

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 $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$.

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

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

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,

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

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

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 :