Machine Learning Theory (CS 6783)

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

1 Recap : Online Learning

For $t = 1$ to n

Instance $x_t \in \mathcal{X}$ is provided

Learner picks $q_t \in \Delta(\mathcal{Y})$

Outcome $y_t \in \mathcal{Y}$ is revealed

Learner draws randomized prediction $\hat{y}_t \sim q_t$ and suffers loss $\ell(\hat{y}_t, y_t)$

end

Goal is to minimize

$$\mathbf{R}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)$$

2 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 ℓ, 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 Φ is some arbitrary function.

3 Minimax Rate for Online Learning

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

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

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 $\langle\!\langle \mathbf{Operator}_t \rangle\!\rangle_{t=1}^n [\dots]$ to refer to $\mathbf{Operator}_1 \mathbf{Operator}_2 \dots \mathbf{Operator}_n [\dots]$. Hence for instance,

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

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] \\
&= \dots \\
&= \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]
\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.

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

where $\mathbf{P}$ *is a joint distribution over the sequence of instances and* $\mathbb{E}_{t-1}\left[\cdot\right]$ *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.

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

□