Machine Learning Theory (CS 6783)

Lecture 17: Minimax Rate for Online Learning

1 Predicting Bit-sequences

Think of the online learning problem where on each round t we we predict the next bit $y_t \in \{\pm 1\}$. Also say $\mathcal{F} \subset \{\pm 1\}^n$ and we want to minimize regret (in expectation) :

$$\mathrm{Reg}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{\hat{y}_t \neq y_t\}} - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}}$$

When can we ensure $\mathbb{E}\left[\mathrm{Reg}_n\right] \to 0$? Let us denote the minimax rate as

$$V_n = \min_{\text{algorithms}} \max_{\text{sequence}} \mathbb{E}\left[\mathrm{Reg}_n\right]$$

Claim 1.

$$V_n = \frac{1}{2n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right]$$

Proof. The basic idea is to write down the minimax rate in a recursive form and get a characterization for it. To this end, say you had already played rounds 1 to $n-1$ optimally, then, on the last two rounds, what are the optimal moves for both the players. We write this value given $y_1, \ldots, y_{n-1}$ were already produced as :

$$V_n(y_1,\ldots,y_{n-1}) = \min_{q_n\in[0,1]} \sup_{y_n\in\{\pm 1\}} \{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\, \mathbb{1}_{\{\hat{y}_n \neq y_n\}}\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t\neq y_t\}}\}$$

That is on the last round, the learner picks distribution q_n that minimizes loss at the last step while the adversary picks y_n that maximizes the loss at last step while also minimizes loss of the target we are comparing our regret against. In fact if we define $V_n(y_1,\ldots,y_n) = -\inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t\neq y_t\}}$ then we see that

$$V_n(y_1,\ldots,y_{n-1}) = \min_{q_n\in[0,1]} \sup_{y_n\in\{\pm 1\}} \{\mathbb{E}_{\hat{y}_n\sim q_n}\left[\, \mathbb{1}_{\{\hat{y}_n \neq y_n\}}\right] + V_n(y_1,\ldots,y_n)\}$$

Thus we see that,

$$\begin{aligned} V_n(y_1,\ldots,y_{n-1}) &= \min_{q_n\in[0,1]} \sup_{y_n\in\{\pm 1\}} \{q_n\, \mathbb{1}_{\{1\neq y_n\}} + (1-q_n)\, \mathbb{1}_{\{1=y_t\}} + V_n(y_1,\ldots,y_n)\} \\ &= \min_{q_n\in[0,1]} \max\left\{(1-q_n) + V_n(y_1,\ldots,y_{n-1},+1), q_n + V_n(y_1,\ldots,y_{n-1},-1)\right\} \end{aligned}$$

Solution is to pick q_n such that the two terms are equal. Hence

$$V_n(y_1,\ldots,y_{n-1}) = \frac{1}{2} + \frac{V_n(y_1,\ldots,y_{n-1},+1) + V_n(y_1,\ldots,y_{n-1},+1)}{2}$$
$$= \frac{1}{2} + \mathbb{E}_{\epsilon_n}\left[V_n(y_1,\ldots,y_{n-1},\epsilon_n)\right]$$

Now recursively we continue as

$$V_n(y_1,\ldots,y_{n-2}) = \min_{q_{n-1}\in[0,1]} \sup_{y_{n-1}\in\{\pm1\}} \left\{q_{n-1}\, \mathbb{1}_{\{1\neq y_n\}} + (1-q_{n-1})\, \mathbb{1}_{\{1=y_t\}} + V_n(y_1,\ldots,y_{n-1})\right\}$$
$$= \frac{1}{2} + \mathbb{E}_{\epsilon_{n-1}}\left[V_n(y_1,\ldots,y_{n-2},\epsilon_{n-1})\right]$$

Proceeding as follows we conclude that :

$$V_n(\cdot) = \mathbb{E}_{\epsilon_1}\left[V_n(\epsilon_1)\right] = \ldots = \mathbb{E}_{\epsilon}\left[V_n(\epsilon_1,\ldots,\epsilon_n)\right]$$

Hence we conclude that :

$$\text{Minimax}_n = \frac{V_n(\cdot)}{n} = \frac{1}{2} + \frac{1}{n}\mathbb{E}_\epsilon\left[-\inf_{f\in\mathcal{F}}\sum_{t=1}^n \mathbb{1}_{\{f_t\neq\epsilon_t\}}\right] = \frac{1}{2} + \frac{1}{2n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n f_t\epsilon_t\right] - \frac{1}{2}$$

□

Prediction algorithm : the prediction algorithm corresponding to the above analysis is exactly the q_t that minimizes the recursion at each step and hence is given by

$$q_t = \operatorname*{argmin}_{q\in[0,1]} \max_{y_t\in\{\pm1\}} \left\{\mathbb{E}_{\hat{y}_t\sim q}\left[\, \mathbb{1}_{\{\hat{y}_t\neq y_t\}}\right] + V_n(y_1,\ldots,y_t)\right\}$$
$$= \frac{1}{2}\left(1 + V_n(y_1,\ldots,y_{t-1},+1) - V_n(y_1,\ldots,y_{t-1},-1)\right)$$
$$= \frac{1}{2}\big(1 + \mathbb{E}_{\epsilon_{t+1:n}}\left[V_n(y_1,\ldots,y_{t-1},+1,\epsilon_{t+1},\ldots,\epsilon_n)\right] - \mathbb{E}_{\epsilon_{t+1:n}}\left[V_n(y_1,\ldots,y_{t-1},-1,\epsilon_{t+1},\ldots,\epsilon_n)\right]\big)$$

In fact, we can also show that the following randomized algorithm works. Draw $\epsilon_{t+1},\ldots,\epsilon_n$ and set :

$$q_t = \tfrac{1}{2}\left(1 + \inf_{f\in\mathcal{F}}\left\{\sum_{j=1}^{t-1} \mathbb{1}_{\{f_t\neq y_t\}} + \mathbb{1}_{\{f_t\neq 1\}} + \sum_{i=t+1}^{n} \mathbb{1}_{\{f_i\neq\epsilon_i\}}\right\} - \inf_{f\in\mathcal{F}}\left\{\sum_{j=1}^{t-1} \mathbb{1}_{\{f_t\neq y_t\}} + \mathbb{1}_{\{f_t\neq -1\}} + \sum_{i=t+1}^{n} \mathbb{1}_{\{f_i\neq\epsilon_i\}}\right\}\right)$$

2 General Online Learning

For a general online learning problem, the minimax rate can be written recursively as:

$$\mathcal{V}_n^{sq}((x_1,y_1),\ldots,(x_n,y_n)) = -\inf_{f\in\mathcal{F}}\sum_{t=1}^n \ell(f(x_t),y_t)$$

and subsequently,

$$\mathcal{V}_n^{sq}((x_1,y_1),\ldots,(x_t,y_t)) = \sup_{x_{t+1}\in\mathcal{X}} \inf_{q_{t+1}\in\Delta(\mathcal{Y})} \sup_{y_{t+1}\in\mathcal{Y}} \left\{\mathbb{E}_{\hat{y}_{t+1}\sim q_{t+1}}\left[\ell(\hat{y}_{t+1},y_{t+1})\right] + \mathcal{V}_n^{sq}((x_1,y_1),\ldots,(x_{t+1},y_{t+1}))\right\}$$

Finally we get

$$\begin{aligned} n\mathcal{V}_n^{sq} = \mathcal{V}_n^{sq}(\cdot) &= \underbrace{\sup_{x_1\in\mathcal{X}} \inf_{q_1\in\Delta(\mathcal{Y})} \sup_{y_1\in\mathcal{Y}} \mathbb{E}_{\hat{y}_1\sim q_1} \ldots}_{\text{repeated}} \sup_{x_n\in\mathcal{X}} \inf_{q_n\in\Delta(\mathcal{F})} \sup_{y_n\in\mathcal{Y}} \mathbb{E}_{\hat{y}_n\sim q_n}\left[\sum_{t=1}^n \ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^n \ell(f(x_t),y_t)\right] \\ &= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{q_t\in\Delta(\mathcal{Y})} \sup_{y_t\in\mathcal{Y}} \mathbb{E}_{\hat{y}_t\sim q_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sum_{t=1}^n \ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^n \ell(f(x_t),y_t)\right] \end{aligned}$$

3 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 2 (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 that is lower semi-continuous. We shall use
$\Phi(y_t) = \mathcal{V}_n^{sq}((x_1,y_1),\ldots,(x_t,y_t))$