CSCI 678: Theoretical Machine Learning
Lecture 5

Fall 2024, Instructor: Haipeng Luo

1 From Statistical Learning to Online Learning

Now that we have developed a quite complete picture for statistical learning, we will move on to the harder online learning setting, which, as mentioned in Lecture 1, completely removes the i.i.d. assumption, captures many modern machine learning applications, and have powerful implications on optimization, game theory, privacy, and other areas. Even though online learning is harder than statistical learning, we will show that one can establish a similar theory on learnability based on similar but slightly more advanced techniques.

First, recall the general setup for online learning, which can be seen as a sequential game between a learner and the environment. The game proceeds in rounds, and for each round $t = 1, \dots, n$, the learner first predicts $\widehat{y}_t \in \mathcal{D}$ while the environment chooses $z_t \in \mathcal{Z}$ simultaneously, then the learner suffers loss $\ell(\widehat{y}_t, z_t)$ and observes z_t. The goal of the learner is to minimize the regret against some reference class $\mathcal{F} \subset \mathcal{D}$,

$$\mathrm{Reg}(\mathcal{F}, n) = \sum_{t=1}^{n} \ell(\widehat{y}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t),$$

and the value of this sequential game can be written as $\mathcal{V}^{\text{seq}}(\mathcal{F}, n) = \inf_\pi \sup_{z_{1:n}} \mathbb{E}\left[\frac{\mathrm{Reg}(\mathcal{F},n)}{n}\right]$, which, as we proved in Lecture 1, is always at least as large as $\mathcal{V}^{\text{iid}}(\mathcal{F}, n)$. $\mathcal{F}$ is said to be online learnable if the value $\mathcal{V}^{\text{seq}}(\mathcal{F}, n)$ goes to 0 as n increases.

Moreover, for an adaptive environment where z_t can depend on $\widehat{y}_1, \dots, \widehat{y}_{t-1}$, the value can be further simplified as

$$\mathcal{V}^{\text{seq}}(\mathcal{F}, n) = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{\mathrm{Reg}(\mathcal{F}, n)}{n}\right].$$

We will focus on adaptive environments (which are harder than oblivious environments) and relax the value $\mathcal{V}^{\text{seq}}(\mathcal{F}, n)$ step by step following the same roadmap for statistical learning.

1.1 Empirical process with dependent data

Recall that in statistical learning, the very first step to relax the value is by choosing a specific learning strategy: ERM, then the value can be shown to be bounded as the expected supremum of an empirical process. Is there a similar analogue for online learning?

The first natural attempt is to do ERM at each step: $\widehat{y}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{\tau=1}^{t-1} \ell(f, z_\tau)$. This is called the *follow-the-leader* approach in online learning, and it turns out that this approach will lead to linear (in n) regret even for very simple problems and is in general not a good algorithm for online learning. We postpone the proof of this claim to future lectures that focus on algorithm design.

So what other algorithms should we try? Instead of searching for different candidates, we will in fact take a bolder approach — directly relax $\mathcal{V}^{\text{seq}}(\mathcal{F}, n)$ *without constructing an algorithm*. This can be done with the help of the celebrated *minimax theorem*. Specifically, we first write the value

as the following equivalent form that introduces randomness to the environment (convince yourself that this is indeed equivalent):

$$\mathcal{V}^{\text{seq}}(\mathcal{F}, n) = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{D})} \sup_{p_t \in \Delta(\mathcal{Z})} \mathbb{E}_{\widehat{y}_t \sim q_t, z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{\text{Reg}(\mathcal{F}, n)}{n} \right].$$

Under some mild technical conditions which hold for all problems we will discuss, minimax theorem says that we can in fact swap all the $\inf$ and $\sup$ above, leading to:

$$\mathcal{V}^{\text{seq}}(\mathcal{F}, n) = \left\langle\!\!\left\langle \sup_{p_t \in \Delta(\mathcal{Z})} \inf_{q_t \in \Delta(\mathcal{D})} \mathbb{E}_{\widehat{y}_t \sim q_t, z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{\text{Reg}(\mathcal{F}, n)}{n} \right].$$

We will not go into the details of these conditions. Instead, let's focus on the consequence of applying minimax theorem above. First, note that we have in some sense swapped the order of the learner and the environment in this sequential game — at each time t, the environment now first comes up with a distribution p_t over the outcome z_t, then the learner, *knowing the distribution* p_t, comes up with a randomized strategy q_t. This is sometimes called the *dual game*. While the dual game is seemingly more favorable for the learner (since they play second now) and might have a smaller value, minimax theorem tells us that in fact the value of the game remains exactly the same! In other words, which player goes first makes no difference as long as both players behave optimally.

Second, note that in the dual game, randomness is not needed for the learner anymore:

$$\mathcal{V}^{\text{seq}}(\mathcal{F}, n) = \left\langle\!\!\left\langle \sup_{p_t \in \Delta(\mathcal{Z})} \inf_{\widehat{y}_t \in \mathcal{D}} \mathbb{E}_{z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{\text{Reg}(\mathcal{F}, n)}{n} \right].$$

This is simply because the best randomized strategy q_t is to put all the mass on the optimal $\widehat{y}_t \in \mathcal{D}$. Note that, however, randomness is required for the environment now. In order words, we have also swapped the randomness in some sense.

Finally, we emphasize that even if one could come up with the exact optimal strategy for the learner in the dual game, it provides no clue on how the learner should behave in the original game (at least not directly), simply because the strategies for these two different games do not even pass "type-checking" — the one in the dual game requires seeing the strategy of the environment first before making its own decision, while the one in the original game needs to make the decision first. Therefore, by going to the dual game, on the one hand we can still argue about the value of the original game, but on the other hand we have in some sense lost all the algorithmic information for the learner. (We will see how to address this in a few weeks though.)

So how is looking at the value of the dual game any easier? It turns out that by only one more step of upper bounding, we can further bound it by the expected supremum of some empirical process *with dependent data*. This is summarized in the following theorem.

Theorem 1. *The value of the dual game is bounded as*

$$\begin{aligned} &\left\langle\!\!\left\langle \sup_{p_t \in \Delta(\mathcal{Z})} \inf_{\widehat{y}_t \in \mathcal{D}} \mathbb{E}_{z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{\text{Reg}(\mathcal{F}, n)}{n} \right] \\ &\le \sup_{\mathcal{P} \in \Delta(\mathcal{Z}^n)} \mathop{\mathbb{E}}_{z_{1:n} \sim \mathcal{P}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}_{z_t' \sim \mathcal{P}(\cdot | z_{1:t-1})} \left[\ell(f, z_t') \right] - \ell(f, z_t) \right) \right]. \end{aligned} \tag{1}$$

To understand this bound, one should compare it with the very similar bound

$$\begin{aligned} \mathcal{V}^{\text{iid}}(\mathcal{F}, n) &\le \sup_{\mathcal{P}} \left(\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left(L(f) - \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right) \right] \right) \\ &= \sup_{\mathcal{P}} \left(\mathop{\mathbb{E}}_{z_{1:n} \sim \mathcal{P}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}_{z_t' \sim \mathcal{P}} \left[\ell(f, z_t') \right] - \ell(f, z_t) \right) \right] \right) \end{aligned} \tag{2}$$

for statistical learning. The two differences are: 1) while $z_1, \ldots, z_n$ are drawn independently from the worst-case distribution $\mathcal{P}$ in Equation (2), they are drawn from a worst-case *joint distribution* $\mathcal{P}$ in Equation (1) and do not need to be independent; 2) in Equation (2), each summand involves a term

$\mathbb{E}\left[\ell(f, z_t')\right]$, which is the expected loss of f under the distribution $\mathcal{P}$ and is the same no matter what t is, while in Equation (1), each summand also involves a term $\mathbb{E}\left[\ell(f, z_t')\right]$, but z_t' is drawn from the conditional distribution of $\mathcal{P}$ given the past $z_{1:t-1}$, and thus is different for different t. Finally, we point out that bound (1) is clearly at least as large as bound (2), since if we restrict $\mathcal{P}$ in Equation (1) to range over product distributions, then the bound becomes exactly the same as Equation (2).

The collection of random variables $\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}_{z_t'\sim P(\cdot|z_{1:t-1})}\left[\ell(f, z_t')\right] - \ell(f, z_t)\right)$ index by $f \in \mathcal{F}$ is called an *empirical process with dependent data*. Note that the conditional expectation of $\mathbb{E}_{z_t'\sim P(\cdot|z_{1:t-1})}\left[\ell(f, z_t')\right] - \ell(f, z_t)$ given $z_{1:t-1}$ is clearly 0 for any t, which means each random variable in this empirical process is in fact the average of a sequence of martingale differences and should be small for each f. Whether the supremum of these random variables is also reasonably small, however, will depend on the structure of $\mathcal{F}$.

Proof of Theorem 1. For simplicity we prove the theorem for $n = 2$. The general case can be proven by following the exact same idea. When $n = 2$, the left hand side multiplied by n is simply

$$\sup_{p_1}\inf_{\widehat{y}_1}\mathbb{E}_{z_1}\left[\sup_{p_2}\inf_{\widehat{y}_2}\mathbb{E}_{z_2}\left[\ell(\widehat{y}_1, z_1) + \ell(\widehat{y}_2, z_2) - \inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right]\right].$$

Paying attention to the dependence of each term, we can rewrite this as (this might look complicated, but note that every step is equality!)

$$\begin{aligned}
&\sup_{p_1}\inf_{\widehat{y}_1}\mathbb{E}_{z_1}\left[\ell(\widehat{y}_1, z_1) + \sup_{p_2}\inf_{\widehat{y}_2}\mathbb{E}_{z_2}\left[\ell(\widehat{y}_2, z_2) - \inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right]\right] \\
&= \sup_{p_1}\left(\inf_{\widehat{y}_1}\mathbb{E}_{z_1'}\left[\ell(\widehat{y}_1, z_1')\right] + \mathbb{E}_{z_1}\sup_{p_2}\inf_{\widehat{y}_2}\mathbb{E}_{z_2}\left[\ell(\widehat{y}_2, z_2) - \inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right]\right) \\
&= \sup_{p_1}\mathbb{E}_{z_1}\sup_{p_2}\left(\inf_{\widehat{y}_1}\mathbb{E}_{z_1'}\left[\ell(\widehat{y}_1, z_1')\right] + \inf_{\widehat{y}_2}\mathbb{E}_{z_2}\left[\ell(\widehat{y}_2, z_2) - \inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right]\right) \\
&= \sup_{p_1}\mathbb{E}_{z_1}\sup_{p_2}\left(\inf_{\widehat{y}_1}\mathbb{E}_{z_1'}\left[\ell(\widehat{y}_1, z_1')\right] + \inf_{\widehat{y}_2}\mathbb{E}_{z_2'}\left[\ell(\widehat{y}_2, z_2')\right] - \mathbb{E}_{z_2}\left[\inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right]\right) \\
&= \sup_{p_1}\mathbb{E}_{z_1}\sup_{p_2}\mathbb{E}_{z_2}\left[\inf_{\widehat{y}_1}\mathbb{E}_{z_1'}\left[\ell(\widehat{y}_1, z_1')\right] + \inf_{\widehat{y}_2}\mathbb{E}_{z_2'}\left[\ell(\widehat{y}_2, z_2')\right] - \inf_{f\in\mathcal{F}}\left(\ell(f, z_1) + \ell(f, z_2)\right)\right] \\
&= \sup_{p_1}\mathbb{E}_{z_1}\sup_{p_2}\mathbb{E}_{z_2}\sup_{f\in\mathcal{F}}\left(\inf_{\widehat{y}_1}\mathbb{E}_{z_1'}\left[\ell(\widehat{y}_1, z_1')\right] + \inf_{\widehat{y}_2}\mathbb{E}_{z_2'}\left[\ell(\widehat{y}_2, z_2')\right] - \ell(f, z_1) - \ell(f, z_2)\right). \qquad (3)
\end{aligned}$$

Here, z_1' and z_2' are random variables drawn from p_1 and p_2 respectively (that is, same as z_1 and z_2). Next, we perform the only upper bounding step — since $\widehat{y}_1$ and $\widehat{y}_2$ are from $\mathcal{D}$, a superset of $\mathcal{F}$, we can replace $\inf_{\widehat{y}_1}$ and $\inf_{\widehat{y}_2}$ by the particular f from the earlier $\sup_{f\in\mathcal{F}}$, arriving at

$$\sup_{p_1}\mathbb{E}_{z_1}\sup_{p_2}\mathbb{E}_{z_2}\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z_1'}\left[\ell(f, z_1')\right] + \mathbb{E}_{z_2'}\left[\ell(f, z_2')\right] - \ell(f, z_1) - \ell(f, z_2).\right)$$

Finally, we look at $\mathbb{E}_{z_1}\sup_{p_2\in\Delta(\mathcal{Z})}$ and note that for each possible draw of z_1, there is a corresponding best distribution p_2. This is the same as swapping the order and let p_2 range over all the mappings from $\mathcal{Z}$ to $\Delta(\mathcal{Z})$: $\sup_{p_2:\mathcal{Z}\to\Delta(\mathcal{Z})}\mathbb{E}_{z_1}$ and let z_2 be drawn from $p_2(\cdot|z_1)$. This implies that the final expression is exactly equal to

$$\sup_{\mathcal{P}\in\Delta(\mathcal{Z}\times\mathcal{Z})}\mathbb{E}_{(z_1,z_2)\sim\mathcal{P}}\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z_1'\sim\mathcal{P}}\left[\ell(f, z_1')\right] + \mathbb{E}_{z_2'\sim\mathcal{P}(\cdot|z_1)}\left[\ell(f, z_2')\right] - \ell(f, z_1) - \ell(f, z_2),\right)$$

which finishes the proof. □

We remark that Equation (3), which is equal to the value of the dual game, reveals that the optimal p_t in fact does not depend on the previous decisions of the learner in the dual game (it does depend on all the previous outcomes $z_{1:t-1}$ though). Similarly, the optimal strategy for the learner at each time t is simply to minimize the expected loss for the current step (knowing the distribution of the current outcome z_t). The intuitive reason for both is that the only dependence of the regret on the learner's decision $\widehat{y}_t$ is through the loss of the current step $\ell(\widehat{y}_t, z_t)$ (in particular, $\widehat{y}_t$ plays no role in the benchmark term in the regret). This simplicity in the structure of the optimal solutions only exists in the dual game, highlighting the importance of applying the minimax theorem to allow us to focus on the dual game.

1.2 Symmetrization and sequential Rademacher complexity

Following the roadmap for statistical learning, the next step is to use symmetrization technique to further relax the expected supremum of the empirical process and arrive at something close to the Rademacher complexity. There are again connections and importance differences between the two settings. One key difference is that we will need the concept of a *$\mathcal{Z}$-valued tree*, which is just a complete binary tree with some value from $\mathcal{Z}$ in each node. More formally, a $\mathcal{Z}$-valued tree $\boldsymbol{z}$ of depth n consists of n mappings $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ where $\boldsymbol{z}_t : \{-1, +1\}^{t-1} \to \mathcal{Z}$ specifies the values of the t-th level of the tree. For a *path* (from the root to a leave of the tree) denoted by $\epsilon_1, \dots, \epsilon_n \in \{-1, +1\}$ (think -1 as left and $+1$ as right), $\boldsymbol{z}_t(\epsilon_{1:t-1})$ for $t = 1, \dots, n$ specify the n values on this path.[1] For notational convenience, we will simply write $\boldsymbol{z}_t(\epsilon_{1:t-1})$ as $\boldsymbol{z}_t(\boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n) \in \{-1, +1\}^n$, even though $\boldsymbol{z}_t$ only takes the first $t-1$ entries of $\boldsymbol{\epsilon}$ as inputs.[2]

With this concept, for any class $\mathcal{H} : \mathcal{Z} \to \mathbb{R}$, we define its *conditional sequential Rademacher complexity* on a given tree $\boldsymbol{z}$ as

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{H}; \boldsymbol{z}) = \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{n} \epsilon_t h(\boldsymbol{z}_t(\boldsymbol{\epsilon})) \right]$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)$ consists of n i.i.d. Rademacher random variables. The (unconditional) sequential Rademacher complexity of $\mathcal{H}$ is defined as

$$\mathcal{R}^{\text{seq}}(\mathcal{H}) = \sup_{\boldsymbol{z}} \widehat{\mathcal{R}}^{\text{seq}}(\mathcal{H}; \boldsymbol{z}) = \frac{1}{n} \sup_{\boldsymbol{z}} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{n} \epsilon_t h(\boldsymbol{z}_t(\boldsymbol{\epsilon})) \right]$$

where $\boldsymbol{z}$ ranges over all possible $\mathcal{Z}$-valued trees of depth n. Compared to the counterparts in the statistical learning setting, the similar part is that we are still basically measuring how well $\mathcal{H}$ can fit random signs, but the key difference is that instead of having n samples $z_{1:n}$, we now have a tree of $2^n - 1$ samples, and the value of the t-th sample depends on the labels for the previous $t-1$ samples $\epsilon_{1:t-1}$. This corresponds to the sequential aspect of the game — the t-th outcome can depend on the entire history prior to round t. Also note that for the (unconditional) sequential Rademacher complexity, we are taking a sup over all the trees, instead of taking an expectation over some distribution over trees. This amounts to the fact that in online learning, there is no distributional assumption on the data.

Now we are ready to state the symmetrization result for online learning.

Theorem 2. *For any joint distribution $\mathcal{P}$, the expected supremum of an empirical process with dependent data drawn from $\mathcal{P}$ is bounded as*

$$\mathop{\mathbb{E}}_{z_{1:n} \sim \mathcal{P}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}_{z_t' \sim P(\cdot | z_{1:t-1})} \left[\ell(f, z_t') \right] - \ell(f, z_t) \right) \right] \le 2\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})),$$

where $\ell(\mathcal{F}) = \{h_f : \mathcal{Z} \to \mathbb{R} \mid f \in \mathcal{F}, h_f(z) = \ell(f, z), \forall z\}$.

Proof. We will again take $n = 2$ as an example to showcase the key idea of the proof, and the general case can be proven in a similar way. We first rewrite the left hand side (multiplied by $n = 2$) as

$$= \mathbb{E}_{z_1 \sim \mathcal{P}, z_2 \sim \mathcal{P}(\cdot | z_1)} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{z_1' \sim \mathcal{P}} \left[\ell(f, z_1') \right] - \ell(f, z_1) + \mathbb{E}_{z_2' \sim \mathcal{P}(\cdot | z_1)} \left[\ell(f, z_2') \right] - \ell(f, z_2) \right).$$

Next we pull the expectations out of the sup and use a similar symmetrization trick to arrive at an upper bound

$$\begin{aligned} &\mathbb{E}_{z_1, z_1' \sim \mathcal{P}, z_2, z_2' \sim \mathcal{P}(\cdot | z_1)} \sup_{f \in \mathcal{F}} \left(\ell(f, z_1') - \ell(f, z_1) + \ell(f, z_2') - \ell(f, z_2) \right) \\ &= \mathbb{E}_{z_1, z_1' \sim \mathcal{P}, z_2, z_2' \sim \mathcal{P}(\cdot | z_1), \epsilon_2} \sup_{f \in \mathcal{F}} \left(\ell(f, z_1') - \ell(f, z_1) + \epsilon_2 (\ell(f, z_2') - \ell(f, z_2)) \right), \end{aligned}$$

[1] Note that a path is actually completely specified by $\epsilon_1, \dots, \epsilon_{n-1}$ already, but we often include ϵ_n as it will be useful for some other purpose.

[2] For all subsequent discussion involving a tree, you should always draw an illustrative picture to help you understand the intuitive idea of different concepts (and there will be many of them coming up).

where ϵ_2 is a Rademacher random variable and the last step holds since z_2 and z_2' are symmetric. Now it is tempting to also introduce another Rademacher random variable ϵ_1 for the part involving z_1 and z_1'. However, directly doing so is in fact *incorrect* and the last expression is *not equal* to the following

$$\mathbb{E}_{z_1,z_1'\sim\mathcal{P},z_2,z_2'\sim\mathcal{P}(\cdot|z_1),\epsilon_{1:2}} \sup_{f\in\mathcal{F}} \left(\epsilon_1(\ell(f,z_1') - \ell(f,z_1)) + \epsilon_2(\ell(f,z_2') - \ell(f,z_2))\right). \qquad (\times)$$

The reason is that z_1 and z_1' are actually not symmetric, since z_2 and z_2' are both drawn from the conditional distribution given z_1, which makes the role of z_1 different from that of z_1'!

To proceed with symmetrization, we will instead have to first remove this extra dependence on z_1 by replacing $\mathbb{E}_{z_2,z_2'}$ with the worst case, leading to an upper bound

$$\mathbb{E}_{z_1,z_1'\sim\mathcal{P}} \sup_{z_2,z_2'} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(\ell(f,z_1') - \ell(f,z_1) + \epsilon_2(\ell(f,z_2') - \ell(f,z_2))\right).$$

Now the role of z_1 and z_1' are exactly the same and we can symmetrize it as

$$\mathbb{E}_{z_1,z_1'\sim\mathcal{P}}\mathbb{E}_{\epsilon_1} \sup_{z_2,z_2'} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(\epsilon_1(\ell(f,z_1') - \ell(f,z_1)) + \epsilon_2(\ell(f,z_2') - \ell(f,z_2))\right),$$

which can be further bounded as

$$\begin{aligned}
&\sup_{z_1,z_1'} \mathbb{E}_{\epsilon_1} \sup_{z_2,z_2'} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(\epsilon_1(\ell(f,z_1') - \ell(f,z_1)) + \epsilon_2(\ell(f,z_2') - \ell(f,z_2))\right) \\
&\le \sup_{z_1,z_1'} \mathbb{E}_{\epsilon_1} \sup_{z_2,z_2'} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(\epsilon_1\ell(f,z_1') + \epsilon_2\ell(f,z_2')\right) + \sup_{z_1,z_1'} \mathbb{E}_{\epsilon_1} \sup_{z_2,z_2'} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(-\epsilon_1\ell(f,z_1) - \epsilon_2\ell(f,z_2)\right) \\
&= 2\sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \sup_{f\in\mathcal{F}} \left(\epsilon_1\ell(f,z_1) + \epsilon_2\ell(f,z_2)\right).
\end{aligned}$$

The final step is similar to the last step of the proof of Theorem 1 — look at $\mathbb{E}_{\epsilon_1}\sup_{z_2}$ and note that for $\epsilon_1 = +1$, there is a corresponding $z_2(+1)$ that "attains" the sup over z_2; and similarly for $\epsilon_1 = -1$, there is a corresponding $z_2(-1)$ that "attains" the sup. Therefore, it makes no difference if we swap $\mathbb{E}_{\epsilon_1}$ and $\sup_{z_2}$, and makes z_2 range over all the possible "level 2" of a tree, leading to

$$2\sup_{\boldsymbol{z}} \mathbb{E}_{\epsilon_{1:2}} \sup_{f\in\mathcal{F}} \left(\epsilon_1\ell(f,z_1) + \epsilon_2\ell(f,z_2)\right).$$

This finishes the proof. □

From the proof, we also see that even if we start from a joint distribution $\mathcal{P}$, because of the step of relaxing $\mathbb{E}$ to sup, we end up having a sup over all the possible trees and lose the information about $\mathcal{P}$ eventually. This is also the reason why sequential Rademacher complexity is defined over the worst-case tree.

1.3 Erasing the loss

Similarly to statistical learning, for many problems, it is possible to ignore the loss function when considering sequential Rademacher complexity, as shown in the following two lemmas.

Lemma 1. *For a binary classification problem with $\mathcal{Z} = \mathcal{X}\times\{-1,+1\}$, $\mathcal{F}\subset\{-1,+1\}^{\mathcal{X}}$, and 0-1 loss, one has for any $\mathcal{Z}$-valued tree $(\boldsymbol{x},\boldsymbol{y})$, there exists another $\mathcal{X}$-valued tree $\boldsymbol{x}'$ such that*

$$\widehat{\mathcal{R}}^{\text{seq}}(\ell(\mathcal{F});(\boldsymbol{x},\boldsymbol{y})) = \frac{1}{2}\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F};\boldsymbol{x}').$$

Therefore we have $\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \le \frac{1}{2}\mathcal{R}^{\text{seq}}(\mathcal{F})$.

Lemma 2 (Contraction)**.** *For a regression problem with $\mathcal{Z} = \mathcal{X}\times\mathbb{R}$ and loss $\ell(f,(x,y)) = \ell'(f(x),y)$ for some loss $\ell'(y',y)$ that is G-Lipschitz in the first parameter, one has*

$$\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \le G\mathcal{R}^{\text{seq}}(\mathcal{F}) \times \mathcal{O}(\ln^{3/2} n).$$

These lemmas are analogues of those in Lecture 2 for statistical learning, with the following differences. For Lemma 1, the statistical learning analogue is $\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F});(x_{1:n},y_{1:n})) = \frac{1}{2}\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F};x_{1:n})$

for any sequence $(x_{1:n}, y_{1:n})$, while for online learning we have moved from a tree $(\boldsymbol{x}, \boldsymbol{y})$ to some other tree $\boldsymbol{x}'$ (the reason will be clearly shown in the proof). Nevertheless, note that this does not affect the final conclusion $\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \le \frac{1}{2}\mathcal{R}^{\text{seq}}(\mathcal{F})$, similar to $\mathcal{R}^{\text{iid}}(\ell(\mathcal{F})) = \frac{1}{2}\mathcal{R}^{\text{iid}}(\mathcal{F})$. For Lemma 2, the same subtly exists, and in addition, we lose a factor of $\mathcal{O}(\ln^{3/2} n)$ compared to the statistical learning analogue. It is not clear if this extra factor is necessary or not.

We omit the proof for Lemma 2 and prove Lemma 1 below.

Proof of Lemma 1. By definition we have

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{seq}}(\ell(\mathcal{F}); (\boldsymbol{x}, \boldsymbol{y})) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \mathbf{1}\left\{f(\boldsymbol{x}_t(\boldsymbol{\epsilon})) \neq \boldsymbol{y}_t(\boldsymbol{\epsilon})\right\}\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t \frac{1 - \boldsymbol{y}_t(\boldsymbol{\epsilon}) f(\boldsymbol{x}_t(\boldsymbol{\epsilon}))}{2}\right] \\
&= \frac{1}{2n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} -\epsilon_t \boldsymbol{y}_t(\boldsymbol{\epsilon}) f(\boldsymbol{x}_t(\boldsymbol{\epsilon}))\right].
\end{aligned}$$

We now claim that the random variables $s_t = -\epsilon_t \boldsymbol{y}_t(\boldsymbol{\epsilon})$ for $t = 1, \dots, n$ are in fact also n i.i.d. Rademacher random variables, or equivalently, the mapping $\boldsymbol{\epsilon} \to \boldsymbol{s} = (s_1, \cdots, s_n)$ is a bijection between $\{-1, +1\}^n$ and itself. Indeed, this is clear by constructing the inverse mapping $\boldsymbol{s} \to \boldsymbol{\epsilon}$ defined by $\epsilon_t = -s_t \boldsymbol{y}_t(\epsilon_{1:t-1})$ (note that $\epsilon_{1:t-1}$ can be further expressed in terms of $\boldsymbol{s}$ recursively).

Based on this fact, we can construct a tree $\boldsymbol{x}'$ such that $\boldsymbol{x}'_t(\boldsymbol{s}) = \boldsymbol{x}_t(\boldsymbol{\epsilon})$ for any $\boldsymbol{\epsilon}$ and t (note that the tree is well defined due to the bijection), and thus

$$\widehat{\mathcal{R}}^{\text{seq}}(\ell(\mathcal{F}); (\boldsymbol{x}, \boldsymbol{y})) = \frac{1}{2n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} s_t f(\boldsymbol{x}'_t(\boldsymbol{s}))\right] = \frac{1}{2}\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}; \boldsymbol{x}').$$

Taking sup over $(\boldsymbol{x}, \boldsymbol{y})$ on both sides further proves $\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \le \frac{1}{2}\mathcal{R}^{\text{seq}}(\mathcal{F})$. □

We remark that the tree $\boldsymbol{x}'$ is constructed by permuting the paths of $\boldsymbol{x}$ according to $\boldsymbol{y}$ in some complicated way. As an illustration, consider $\boldsymbol{y}$ being the tree with $+1$ in all nodes. Then it is not hard to see that $\boldsymbol{x}'$ is exactly the mirror reflection of $\boldsymbol{x}$. As another example, if $\boldsymbol{y}$ has $+1$ in the root and -1 everywhere else, then $\boldsymbol{x}'$ is obtained by swapping the left and right subtrees of the root of $\boldsymbol{x}$.

2 Finite class

From now on we will focus on bounding $\mathcal{R}^{\text{seq}}(\mathcal{F})$ for some function class $\mathcal{F}$, starting with a finite class in this section. The key is to apply maximal inequality again, restated below for convenience.

Lemma 3 (Maximal Inequality)**.** *Suppose $\{U_f\}_{f\in\mathcal{F}}$ is a finite collection of σ-sub-Gaussian random variables. Then we have*

$$\mathbb{E}\left[\max_{f\in\mathcal{F}} U_f\right] \le \sigma\sqrt{2\ln|\mathcal{F}|}.$$

The main result is stated below.

Theorem 3. *Let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class. We have for any $\mathcal{X}$-valued tree $\boldsymbol{x}$,*

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}; \boldsymbol{x}) \le \frac{1}{n}\sqrt{2\left(\max_{f\in\mathcal{F}}\max_{\boldsymbol{\epsilon}} \sum_{t=1}^{n} f^2(\boldsymbol{x}_t(\boldsymbol{\epsilon}))\right)\ln|\mathcal{F}|}.$$

Consequently, if $\mathcal{Y} \subset [-C, C]$ for some $C > 0$, then $\mathcal{R}^{\text{seq}}(\mathcal{F}) \le C\sqrt{\frac{2\ln|\mathcal{F}|}{n}}$.

Proof. Note that $\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}; \boldsymbol{x}) = \frac{1}{n}\mathbb{E}\left[\max_{f\in\mathcal{F}} U_f\right]$ where $U_f = \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t(\boldsymbol{\epsilon}))$. Below we show that U_f is σ-sub-Gaussian with $\sigma = \max_{f\in\mathcal{F}}\max_{\boldsymbol{\epsilon}}\sqrt{\sum_{t=1}^{n} f^2(\boldsymbol{x}_t(\boldsymbol{\epsilon}))}$, so applying maximal inequality then finishes the proof.

Indeed, with $U_{f,\tau} = \sum_{t=1}^{\tau} \epsilon_t f(\boldsymbol{x}_t(\boldsymbol{\epsilon}))$ we have for any $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda U_{f,n}\right)\right] &= \mathbb{E}\left[\exp\left(\lambda U_{f,n-1}\right)\mathbb{E}\left[\exp\left(\lambda\epsilon_n f(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right) \mid \epsilon_{1:n-1}\right]\right] \\ &\le \mathbb{E}\left[\exp\left(\lambda U_{f,n-1}\right)\exp\left(\tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right)\right]\end{aligned}$$

where the inequality is by the fact that $\epsilon_n f(\boldsymbol{x}_n(\boldsymbol{\epsilon}))$ is $|f(\boldsymbol{x}_n(\boldsymbol{\epsilon}))|$-sub-Gaussian. Continuing to peel the last term of $U_{f,n-1}$ in the same way, we arrive at

$$\mathbb{E}\left[\exp\left(\lambda U_{f,n-2}\right)\mathbb{E}\left[\exp\left(\lambda\epsilon_{n-1} f(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon}))\right)\exp\left(\tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right) \mid \epsilon_{1:n-2}\right]\right],$$

but note that the term $\exp\left(\frac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right)$ also involves the randomness of ϵ_{n-1}, so we cannot directly proceed in the same way. Instead, we bound it by considering the worst case:

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda U_{f,n-2}\right)\mathbb{E}\left[\exp\left(\lambda\epsilon_{n-1} f(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon}))\right) \mid \epsilon_{1:n-2}\right]\max_{\epsilon_{n-1}}\exp\left(\tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right)\right] \\ &\le \mathbb{E}\left[\exp\left(\lambda U_{f,n-2}\right)\max_{\epsilon_{n-1}}\exp\left(\tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon})) + \tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right)\right] \\ &\qquad\qquad (\epsilon_{n-1} f(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon})) \text{ is } |f(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon}))|\text{-sub-Gaussian}) \\ &\le \mathbb{E}\left[\exp\left(\lambda U_{f,n-2}\right)\max_{\epsilon_{n-2},\epsilon_{n-1}}\exp\left(\tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_{n-1}(\boldsymbol{\epsilon})) + \tfrac{1}{2}\lambda^2 f^2(\boldsymbol{x}_n(\boldsymbol{\epsilon}))\right).\right]\end{aligned}$$

Continuing in the same fashion, we arrive at

$$\mathbb{E}\left[\exp\left(\lambda U_{f,n}\right)\right] \le \max_{\boldsymbol{\epsilon}}\exp\left(\frac{\lambda^2}{2}\sum_{t=1}^{n} f^2(\boldsymbol{x}_t(\boldsymbol{\epsilon}))\right) \le \exp\left(\lambda^2\sigma^2/2\right),$$

which shows that U_f is σ-sub-Gaussian. □

This shows that any finite class with bounded value is online learnable, which will play a key role in following discussion with infinite classes.

3 Infinite Class: Online Binary Classification

Next, we move on to discuss the learnability of infinite classes, starting from binary classification with 0-1 loss. Recall that for statistical learning, we made a key observation that even if $\mathcal{F}$ is infinite, what really matters is the projection $\mathcal{F}|_{x_{1:n}}$, which is always finite. Similarly, for online learning we also have

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F};\boldsymbol{x}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\boldsymbol{x}_t(\boldsymbol{\epsilon}))\right] = \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{v}\in V}\sum_{t=1}^{n}\epsilon_t \boldsymbol{v}_t(\boldsymbol{\epsilon})\right] \le \sqrt{\frac{2\ln|V|}{n}} \tag{4}$$

where $V = \mathcal{F}|_{\boldsymbol{x}} = \{(f\circ\boldsymbol{x}_1,\cdots,f\circ\boldsymbol{x}_n) \mid f\in\mathcal{F}\}$ is the projection of $\mathcal{F}$ onto tree $\boldsymbol{x}$, which is a set of $\{-1,+1\}$-valued trees. Note that $\mathcal{F}|_{\boldsymbol{x}}$ is always finite, so we have yet again moved from an infinite class to a finite class. However, how large can $|\mathcal{F}|_{\boldsymbol{x}}|$ be? Since a tree of depth n has 2^n-1 nodes, the cardinality of $\mathcal{F}|_{\boldsymbol{x}}$ can be as bad as 2^{2^n-1}, leading to a vacuous bound. On the other hand, recall that in statistical learning, for a set of n samples $x_{1:n}$, $|\mathcal{F}|_{x_{1:n}}|$ can only be at most 2^n.

Since both 2^{2^n-1} and 2^n are vacuous bounds anyway, maybe we should just hope that $|\mathcal{F}|_{\boldsymbol{x}}|$ is small for common problems with a class $\mathcal{F}$ that is not too complex? This is unfortunately not true, since $|\mathcal{F}|_{\boldsymbol{x}}|$ can be way too large even for a very simple class. To see this, consider the following class defined over $\mathcal{X} = \mathbb{R}$:

$$\mathcal{F} = \left\{ f_\theta(x) = \begin{cases} +1, & \text{if } x = \theta \\ -1, & \text{else} \end{cases} \;\middle|\; \theta\in\mathbb{R} \right\}. \tag{5}$$

This class is intuitively simple since each classifier f_θ in the class is predicting $+1$ for one and only one specific input θ. Indeed, it is clear that this class cannot even shatter a set of size two, and thus $\text{VCdim}(\mathcal{F}) = 1$, which means it is (easily) learnable in the statistical learning setting.

However, it is easy to construct a tree such that $|\mathcal{F}|_{\boldsymbol{x}}| = 2^n$, which again makes the bound in Equation (4) vacuous. To show this, simply let $\boldsymbol{x}$ have distinct values in all nodes. Then $\mathcal{F}|_{\boldsymbol{x}}$

contains the tree that has -1 in all nodes, and $2^n - 1$ other different trees, each of which has one and only one node with value $+1$.

So does this mean that $|\mathcal{F}|_{\boldsymbol{x}}|$ is not the right complexity measure, or is this simple class really not online learnable? It would be very unfortunate if even a class as simple as this is not online learnable. Fortunately, it turns out that this is not the case and the projection is really not the right concept to consider. To see how to fix this, note that the projection is really a set V of $\{-1, +1\}$-valued trees, such that

$$\forall f \in \mathcal{F}, \exists \boldsymbol{v} \in V, \text{s.t. } \forall \boldsymbol{\epsilon} \in \{-1, +1\}^n, f(\boldsymbol{x}_t(\boldsymbol{\epsilon})) = \boldsymbol{v}_t(\boldsymbol{\epsilon}) \text{ holds for all } t = 1, \dots, n.$$

However, suppose that we have a set V of $\{-1, +1\}$-valued trees such that a similar statement holds but importantly with two quantifiers swapped:

$$\forall f \in \mathcal{F}, \forall \boldsymbol{\epsilon} \in \{-1, +1\}^n, \exists \boldsymbol{v} \in V, \text{s.t. } f(\boldsymbol{x}_t(\boldsymbol{\epsilon})) = \boldsymbol{v}_t(\boldsymbol{\epsilon}) \text{ holds for all } t = 1, \dots, n.$$

Then this is in fact already enough for Equation (4) to hold (try to convince yourself)! A set V with the property above is called an *zero-cover* of $\mathcal{F}|_{\boldsymbol{x}}$, and the zero-covering number $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ is defined as the size of the smallest zero-cover. We have thus shown the following:

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) \leq \sqrt{\frac{2 \ln \mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})}{n}}.$$

So how large can $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ be then? First of all, this is clearly always not larger than $|\mathcal{F}|_{\boldsymbol{x}}|$ (since $\mathcal{F}|_{\boldsymbol{x}}$ is a zero-cover of itself). Second, $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ is in fact always bounded by 2^n. This is because the set of all the possible trees with the same value at each level is always a zero-cover for any class, and there are clearly 2^n such trees (since each level takes one of the two possible values). This is of course still a vacuous bound, but it is at least the same vacuous bound as the one for a projection in statistical learning, indicating that this might be the right complexity measure.

For a class with specific structures, $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ can be much smaller. For example, the simple class defined in Equation (5) has zero-covering number $n + 1$, implying that it is online learnable (as we hope). To get an intuition on why its zero-covering number is $n + 1$, consider a special case when $\boldsymbol{x}$ contains no identical value along any path. Then, we only need the following $n + 1$ trees to cover $\mathcal{F}|_{\boldsymbol{x}}$: a tree with -1 in every node, and for each $t = 1, \dots, n$, a tree with $+1$ for all nodes at level t and -1 everywhere else.

So what about the general case with possible identical values on a path? Directly constructing a cover seems to be challenging this case, and it would be ideal if there exists a combinatorial parameter, similar to VC dimension, that is easier to compute and can be used to provide a good upper bound on the zero-covering number, in the same way as what Sauer's lemma does in the statistical learning. In the next lecture, we will see that there is indeed such an analogue.