Theoretical Machine Learning
Lecture 4

Instructor: Haipeng Luo

1 Pseudo-dimension and Fat-shattering Dimension

In the last lecture we discussed how covering number can be used to measure the complexity of learning a class of real-valued functions, very similar to the role of growth function for classification problems. Recall that for binary classification we also introduced VC dimension, a combinatorial parameter of a class that might be easier to figure out and that gives a direct upper bound on the growth function via Sauer's lemma. This leads to a natural question: can we also come up with some combinatorial parameter for a real-valued function class that helps us bound the covering number directly?

Indeed, such combinatorial parameters exist. The first such one in the literature is the *pseudo-dimension*, and it is based on a pretty natural idea of reducing a real-valued function to a binary classifier by looking at it *epigraph*. Specifically, a function $f : \mathcal{X} \to [-1, +1]$ naturally separates the space $\mathcal{X} \times [-1, +1]$ into two parts: the part where $f(x) \leq y$ (which is called the epigraph of f) and the part where $f(x) > y$. Therefore, we can see f as a binary classifier for the space $\mathcal{X} \times [-1, +1]$. Pseudo-dimension of $\mathcal{F}$ is simply defined as the VC dimension of this induced class of binary classifiers:

$$\text{Pdim}(\mathcal{F}) = \text{VCdim}\left(\{h(x, y) = \text{sign}(f(x) - y) \mid f \in \mathcal{F}\}\right).$$

If we spell out the definition of VC dimension, then Pseudo-dimension is the largest number n such that there exist n input-output pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times [-1, +1]$, such that for any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, there exists $f \in \mathcal{F}$ with $\text{sign}(f(x_t) - y_t) = s_t$ for all $t = 1, \ldots, n$. (Try to draw a picture for the case $\mathcal{X} = \mathbb{R}$ to help understand this.)

For example, in the last lecture we discussed the linear class defined by $\mathcal{X} = B_q^d$, $\mathcal{F} = \left\{f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_p^d\right\}$ for some $p \geq 1$ and $q \geq 1$ such that $1/p + 1/q = 1$. To see how large the pseudo-dimension is for this class, we need to look at the VC dimension of the class $\left\{h(x, y) = \text{sign}(\langle \theta, x \rangle - y) \mid \theta \in B_p^d\right\}$. This is very similar to the class of linear classifiers we discussed in Lecture 2 (and HW 1), and it is not hard to verify that the VC dimension is exactly d. Therefore, the pseudo-dimension of $\mathcal{F}$ is d.

A finite pseudo-dimension turns out to be sufficient for learning. Indeed, one can show an analogue of Sauer's lemma which says that the log α-covering number $\ln \mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha)$ is of order $\text{Pdim}(\mathcal{F}) \ln\left(\frac{1}{\alpha}\right)$ (ignoring some log factors). We will not prove this fact, but using this bound with Theorem 2 of Lecture 3 directly gives $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}(\sqrt{\text{Pdim}(\mathcal{F})(\ln n)/n})$. Also note that for the linear class, this gives almost the same bound as what we proved last time.

However, it turns out that finite pseudo-dimension is *not necessary* for learning. To see this, we examine the class of all non-decreasing functions again. The claim is that while this class is learnable (as we proved last time), it actually has infinite pseudo-dimension, which implies that pseudo-dimension is not the "right" complexity measure. Indeed, for any n, consider the input-output pairs $(0, 0/n), (1, 1/n), (2, 2/n), \ldots$. For any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, we can always find a non-decreasing function that passes through the points $(0, 0/n + s_1\epsilon), (1, 1/n + s_2\epsilon), (2, 2/n + s_3\epsilon), \ldots$, as long as ϵ is in $(0, \frac{1}{2n}]$, and it is clear that such a function satisfies $\text{sign}(f(x_t) - y_t) = s_t$ for all

$t = 1, \ldots, n$. This shows that the induced binary classifier class can shatter this kind of training set for any n, and thus the pseudo-dimension is infinity.

If we compare the definition of pseudo-dimension and covering number, what is missing for pseudo-dimension is the "scale" α. Intuitively, we also want to come up with a combinatorial parameter in terms of some scale α, such that it becomes smaller when the scale is larger. One way to do this is to require the induced binary classifier class to not only predict correctly the labels, but also predict correctly with a certain confidence or margin. This leads to the concept of *fat-shattering*. We say that a class $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ α-shatters a set $x_1, \ldots, x_n \in \mathcal{X}$, if there exist $y_1, \ldots, y_n \in [-1, +1]$ (called the *witness to shattering*), such that for any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, there exists $f \in \mathcal{F}$ with $s_t(f(x_t) - y_t) \geq \alpha/2$ for all $t = 1, \ldots, n$. The condition $s_t(f(x_t) - y_t) \geq \alpha/2$ exactly corresponds to predicting the label s_t correctly with margin $\alpha/2$. With this concept, the *fat-shattering dimension* of $\mathcal{F}$ at scale α is defined as the size of the largest α-shattered set:

$$\text{fat}(\mathcal{F}, \alpha) = \max\left\{n \mid \text{there exists a set } x_{1:n} \text{ } \alpha\text{-shattered by } \mathcal{F}\right\}.$$

Clearly, fat$(\mathcal{F}, \alpha)$ is decreasing in α — if $\mathcal{F}$ α-shatters a set, then it must α'-shatters the same set for any $\alpha' < \alpha$ by definition. It is also clear that when α goes to zero, fat-shattering dimension just becomes pseudo-dimension.

Coming back to the example of the class of all non-decreasing functions, we see that in the previous construction of the shattered set, the margin is only $\epsilon \in (0, \frac{1}{2n}]$, which becomes smaller and smaller as we increase n. So if we require the margin to be at least $\alpha/2$ for some α, then the construction will fail for n larger than $1/\alpha$. Indeed, one can prove that the fat-shattering dimension is exactly of order $1/\alpha$.

Proposition 1. *If $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [-1, +1]$, and $\mathcal{F}$ is the set of all non-decreasing functions, then* fat$(\mathcal{F}, \alpha) \leq \frac{4}{\alpha}$ *for any $\alpha \leq 4$.*

Proof. Suppose $x_1 \leq \cdots \leq n$ is α-shattered by $\mathcal{F}$ with witness $y_1 \leq \cdots \leq y_n$. Let $s_t = +1$ for every odd t and $s_t = -1$ for every even t, and $f \in \mathcal{F}$ be the corresponding function that predicts these labels correctly with margin $\alpha/2$. Then by the fact that f is non-decreasing, for every odd t we must have

$$y_{t+1} - y_t \geq y_{t+1} - f(x_{t+1}) + f(x_t) - y_t \geq s_{t+1}(f(x_{t+1}) - y_{t+1}) + s_t(f(x_t) - y_t) \geq \alpha.$$

Since all y_t's are in $[-1, +1]$, we conclude that n must not be larger than $4/\alpha$, finishing the proof. □

So how is the fat-shattering dimension connected to the covering number? It turns out there is also a similar analogue to Sauer's lemma, which we state below without going into the proof.

Theorem 1. *For any $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ and $\alpha \in (0, 1)$, we have for any inputs $x_{1:n}$,*

$$\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha) = \mathcal{O}\left(\text{fat}(\mathcal{F}, c\alpha) \ln\left(\frac{1}{\alpha}\right)\right)$$

for some absolute constant $c > 0$.

The bound is tighter than the one for pseudo-dimension since fat$(\mathcal{F}, c\alpha) \leq$ Pdim$(\mathcal{F})$. Applying Dudley integral entropy further gives us a bound on the Rademacher complexity. For example, applying it to the class of non-decreasing functions gives the following:

Proposition 2. *If $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [-1, +1]$, and $\mathcal{F}$ is the set of all non-decreasing functions, then $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}(\sqrt{1/n})$.*

Proof. We apply Dudley integral entropy with $\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha) = \mathcal{O}\left(\frac{1}{\alpha} \ln\left(\frac{1}{\alpha}\right)\right) = \mathcal{O}\left(\frac{1}{\alpha^{3/2}}\right)$:

$$\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}\left(\inf_{\alpha}\left(\alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{1} \frac{d\delta}{\delta^{3/4}}\right)\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

□

Compared to the bound $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}(\sqrt{(\ln n)/n})$ we obtained in the last lecture via a bound $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha) \leq (n+1)^{\frac{1}{\alpha}}$ on the ℓ_∞ covering number, here we further improve it (removing the $\ln n$ factor) by using a direct bound on the ℓ_2 covering number via fat-shattering dimension. This shows the advantage of going for the fat-shattering dimension directly.

Summary. To summarize, for real-valued function class we have obtained a sequence of upper bounds on the value of the statistical learning game:

$$\begin{aligned}
\mathcal{V}^{\text{iid}}(\mathcal{F},n) &\le \sup_{\mathcal{P}}\left(\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f,z_t)\right)\right]\right) && \text{(using ERM)}\\
&\le 2\sup_{\mathcal{P}}\mathcal{R}^{\text{iid}}(\ell(\mathcal{F})) && \text{(symmetrization)}\\
&\le 2G\sup_{\mathcal{P}}\mathcal{R}^{\text{iid}}(\mathcal{F}) && \text{(erasing the loss)}\\
&\le 2G\sup_{x_{1:n}}\min_{0\le\alpha\le 1}\left(4\alpha+\frac{12}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\ln\mathcal{N}_2(\mathcal{F}|_{x_{1:n}},\delta)}d\delta\right) && \text{(Dudley entropy integral)}\\
&\le 2G\min_{0\le\alpha\le 1}\mathcal{O}\left(\alpha+\frac{1}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\text{fat}(\mathcal{F},c\delta)\ln\left(\frac{1}{\delta}\right)}d\delta\right).
\end{aligned}$$

As for classification problems, we remark that these upper bounds are also pretty tight, in the sense that one can show that finite fat-shattering dimension is *necessary* for the learnability of $\mathcal{F}$. This concludes all the topics for statistical learning covered in this course (we will come back for more in the final project though).

2 From Statistical Learning to Online Learning

After developing a rather complete picture for the learnability of statistical learning, next we will move on to the harder online learning setting and establish a similar theory 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 $\mathcal{F} \subset \mathcal{D}$,

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

and the value of this sequential game can be written as $\mathcal{V}^{\text{seq}}(\mathcal{F},n) = \inf_{\pi}\sup_{z_{1:n}}\mathbb{E}\left[\frac{\text{Reg}(\mathcal{F},n)}{n}\right]$, 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

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

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.

2.1 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 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: $\widehat{y}_t = \text{argmin}_{f\in\mathcal{F}}\sum_{\tau=1}^{t-1}\ell(f,z_\tau)$. This is called the *follow-the-leader* approach in online learning, and it turns out that this approach will lead to linear regret (linear in n) even for very simple problems and is in general not a good algorithm for online learning. We postpone the proof to the second half of the course that focuses 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 randomize the decisions of the environment:

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

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, but will come back to the proof of some version of the minimax theorem later in this course. 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 seemingly the dual game is more favorable for the learner (since he/she plays 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 worst-case $\widehat{y}_t \in \mathcal{D}$. 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 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 2. *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 will depend on the structure of $\mathcal{F}$.

Proof of Theorem 2. 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!)

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

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 decisions of the learner in the dual game (it does depend on all the previous outcomes $z_{1:t-1}$ though). This is the key advantage of going to the dual game and it allows us to simplify the bound greatly.

2.2 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 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ 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* of length n denoted by $\epsilon_1, \ldots, \epsilon_n \in \{-1,+1\}$ (think -1 as left and $+1$ as right), $\boldsymbol{z}_t(\epsilon_{1:t-1})$ for $t = 1, \ldots, n$ specify the n values on this path. 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.

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 3. *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

$$\begin{aligned}
&\mathbb{E}_{z_1,z'_1\sim\mathcal{P}, z_2,z'_2\sim\mathcal{P}(\cdot|z_1)} \sup_{f\in\mathcal{F}}\left(\ell(f,z'_1) - \ell(f,z_1) + \ell(f,z'_2) - \ell(f,z_2)\right) \\
&= \mathbb{E}_{z_1,z'_1\sim\mathcal{P}, z_2,z'_2\sim\mathcal{P}(\cdot|z_1),\epsilon_2} \sup_{f\in\mathcal{F}}\left(\ell(f,z'_1) - \ell(f,z_1) + \epsilon_2(\ell(f,z'_2) - \ell(f,z_2))\right),
\end{aligned}$$

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 2 — look at $\mathbb{E}_{\epsilon_1}\sup_{z_2}$ and note that for $\epsilon = +1$, there is a corresponding $z_2(+1)$ that "attains" the $\sup$ over z_2; and similarly for $\epsilon = -1$, 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.

To sum up, we have successfully derived the first two steps of relaxation for the value of an online learning game, similar to statistical learning:

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

In the next lecture, we will continue this roadmap to further simplify the bound.