Theoretical Machine Learning
Lecture 5

Instructor: Haipeng Luo

1 Erasing the loss and Finite Class Results

Recall that in the last lecture, following the roadmap for statistical learning, we derived the first two steps of relaxation for the value of online learning and arrived at an upper bound in terms of the sequential Rademacher complexity of $\ell(\mathcal{F})$:

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

Next, we further relate it to the sequential Rademacher complexity of $\mathcal{F}$ when $\mathcal{F}$ is a function class, and also derive a bound for the finite case.

1.1 Erasing the loss

For many problems (such as online supervised learning), we have $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$. Similarly to statistical learning, below we show that in this case we can simply ignore the loss when discussing the learnability of the problem.

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 $\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \leq \frac{1}{2}\mathcal{R}^{\text{seq}}(\mathcal{F})$.

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*

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

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 $\mathcal{R}^{\text{seq}}(\ell(\mathcal{F})) \leq \frac{1}{2}\mathcal{R}^{\text{seq}}(\mathcal{F})$, 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 $\boldsymbol{x}_t(\boldsymbol{\epsilon}) = \boldsymbol{x}'_t(\boldsymbol{s})$ 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 $\boldsymbol{x}$.

1.2 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. 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 1. *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 the worst case:

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

Continuing in the same fashion, we arrive at

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

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

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

2 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

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

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 very trivial 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 trivial 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 $\mathcal{X}=\mathbb{R}$:

$$\mathcal{F}=\left\{f_\theta(x)=\begin{cases}+1, & \text{if } x=\theta\\ -1, & \text{else}\end{cases}\;\middle|\;\theta\in\mathbb{R}\right\}.\tag{2}$$

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 $|\mathcal{F}|_{\boldsymbol{x}}|=2^{n-1}$, which again makes the bound in Equation (1) trivial. To show this, simply let $\boldsymbol{x}$ have distinct values in all the leaves. Then $\mathcal{F}|_{\boldsymbol{x}}$ at least contains 2^{n-1} different trees, each of which has a different leaf 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

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

However, suppose 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 (1) to hold (try to convince yourself)! A set V with the above property 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:

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

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 trivial bound, but it is at least the same trivial 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 (2) has zero-covering number $n+1$, implying that it is online learnable (as we hope). We defer the formal proof to the next subsection, but illustrate with a simpler case when $\boldsymbol{x}$ contains no identical value along any path. In this case 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.

2.1 Combinatorial parameter

Similarly to previous discussions for statistical learning, next we will introduce some combinatorial parameter that provides a reasonable upper bound for $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ and that is easier to bound. First, we say that $\mathcal{F}$ shatters a tree $\boldsymbol{x}$ if for any $\boldsymbol{\epsilon} \in \{-1,+1\}^n$, there exists $f \in \mathcal{F}$ such that $f(\boldsymbol{x}_t(\boldsymbol{\epsilon})) = \epsilon_t$ holds for all $t = 1, \dots, n$. Now, we define the *Littlestone* dimension $\text{Ldim}(\mathcal{F})$ of a class $\mathcal{F}$ as the depth of the largest tree that can be shattered by $\mathcal{F}$ ($\text{Ldim}(\mathcal{F})$ is defined as 0 if no tree can be shattered by $\mathcal{F}$, and ∞ if for any n there exists a tree of depth n that is shattered by $\mathcal{F}$).

Note that one always has $\text{VCdim}(\mathcal{F}) \le \text{Ldim}(\mathcal{F})$, since if $x_{1:n}$ is shattered by $\mathcal{F}$, then a tree $\boldsymbol{x}$ of depth n such that all nodes at level t have value x_t for $t = 1, \dots, n$ is clearly shattered by $\mathcal{F}$ as well. Using this fact, one sees that $\text{Ldim}(\mathcal{F}) = 0$ implies $\text{VCdim}(\mathcal{F}) = 0$ and thus $\mathcal{F}$ contains only one function.

Similar to VC dimension, to prove $\text{Ldim}(\mathcal{F}) = d$ we have to 1) construct a tree of depth d that can be shattered by $\mathcal{F}$ and 2) prove that no tree of depth $n+1$ can be shattered by $\mathcal{F}$. As an example, we argue that the simple class defined by Equation (2) has Littlestone dimension exactly 1 (the same as its VC dimension). Clearly it can shatter a tree with depth 1 (in fact, any tree with depth 1). On the other hand, it cannot shatter any tree with depth 2 (just consider $\boldsymbol{\epsilon} = (+1,+1)$ and $\boldsymbol{\epsilon} = (+1,-1)$).

The following theorem is an exact analogue of Sauer's lemma and it provides a bound on the zero-covering number in terms of the Littlestone dimension. Its proof also reveals how to construct a zero-cover recursively.

Theorem 2. *Suppose $\boldsymbol{x}$ is any $\mathcal{X}$-valued tree with depth n and $\mathcal{F} \subset \{-1,+1\}^{\mathcal{X}}$ has Littlestone dimension $d \le n$, then*

$$\mathcal{N}_0(F|_{\boldsymbol{x}}) \le \sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{en}{d}\right)^d.$$

Proof. Let $g(d,n) = \sum_{i=0}^{d} \binom{n}{i}$. We will prove $\mathcal{N}_0(F|_{\boldsymbol{x}}) \le g(d,n)$ via induction on the value of $d+n$ (the second inequality has been proven in Lecture 2). The base case $d+n=1$ is trivial since the only configuration is $d=0$ and $n=1$, in which case $\mathcal{N}_0(F|_{\boldsymbol{x}}) = 1$ clearly. Now we assume that the statement holds for any $n' > d'$ such that $n' + d' < n + d$, and prove $\mathcal{N}_0(F|_{\boldsymbol{x}}) \le g(d,n)$. The case when $d = 0$ is again trivial, so we assume $d > 0$.

First, we provide a way to recursively construct a zero-cover for $\mathcal{F}|_{\boldsymbol{x}}$. Depending on the prediction for the root of $\boldsymbol{x}$, we split the class $\mathcal{F}$ into two subclasses: $\mathcal{F}_- = \{f \in \mathcal{F} \mid f(\boldsymbol{x}_1) = -1\}$ and

$\mathcal{F}_+ = \{f \in \mathcal{F} \mid f(\boldsymbol{x}_1) = -1\}$. Let $\boldsymbol{x}^\ell$ and $\boldsymbol{x}^r$ be the left and right subtrees of the root of $\boldsymbol{x}$ (which have depth $n-1$), and V_+^ℓ and V_+^r be the smallest zero-cover of $\mathcal{F}_+|_{\boldsymbol{x}^\ell}$ and $\mathcal{F}_+|_{\boldsymbol{x}^r}$ respectively. Now we construct a set V_+ in the following way: 1) the root of every tree in V_+ has value $+1$, 2) pair the element from V_+^ℓ and V_+^r to form the left and right subtrees of the root for trees in V_+, so that each element from V_+^ℓ and V_+^r appears at least once. It is clear that this can be done such that $|V_+| = \max\left\{|V_+^\ell|, |V_+^r|\right\}$. Moreover, it is also clear that V_+ is a zero-cover of $\mathcal{F}_+|_{\boldsymbol{x}}$. In the exact same way, we construct V_- such that $|V_-| = \max\left\{|V_-^\ell|, |V_-^r|\right\}$ and V_- is a zero-cover of $\mathcal{F}_-|_{\boldsymbol{x}}$. Finally, we have that $V_- \cup V_+$ is a zero-cover of $\mathcal{F}|_{\boldsymbol{x}}$.

It remains to bound $|V_-|$ and $|V_+|$. The key observation is that it is impossible that $\mathcal{F}_-$ and $\mathcal{F}_+$ both have Littlestone dimension d. Otherwise, there are trees $\boldsymbol{x}^-$ and $\boldsymbol{x}^+$ of depth n that can be shattered by $\mathcal{F}_-$ and $\mathcal{F}_+$ respectively. By pairing $\boldsymbol{x}^-$ and $\boldsymbol{x}^+$ as the left and right subtrees of the root $\boldsymbol{x}_1$, we obtain a tree with depth $n+1$ that can be shattered by $\mathcal{F}$, which is a contradiction to $\mathrm{Ldim}(\mathcal{F}) = d$. Without loss of generality, we can thus assume $\mathcal{F}_-$ has Littlestone dimension at most $d-1$. Using the inductive hypothesis, we thus have

$$|V_+| = \max\left\{|V_+^\ell|, |V_+^r|\right\} = \max\left\{\mathcal{N}_0(\mathcal{F}_+|_{\boldsymbol{x}^\ell}), \mathcal{N}_0(\mathcal{F}_+|_{\boldsymbol{x}^r})\right\} \le g(d, n-1),$$

and

$$|V_-| = \max\left\{|V_-^\ell|, |V_-^r|\right\} = \max\left\{\mathcal{N}_0(\mathcal{F}_-|_{\boldsymbol{x}^\ell}), \mathcal{N}_0(\mathcal{F}_-|_{\boldsymbol{x}^r})\right\} \le g(d-1, n-1).$$

Therefore, $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \le |V| = |V_-| + |V_+| \le g(d-1, n-1) + g(d, n-1) = g(d, n)$ (the last step is proven in Lecture 2). This finishes the proof. □

We remark that the concept of zero-covering is the key to allow a construction with $|V_+| = \max\left\{|V_+^\ell|, |V_+^r|\right\}$. If we focus on the projection instead, we would have arrived at something like $|V_+| = |V_+^\ell| \times |V_+^r|$. Applying this theorem directly, we conclude that the zero-covering number of the simple class defined by Equation (2) is indeed bounded by $g(1, n) = n + 1$ (even if the tree contains identical elements in some paths). The proof of Theorem 2 also reveals a recursive way to construct such a cover with size $n+1$ (try to construct explicitly a zero-cover for a small tree of depth say 3).

The following bound on the value of the game is a direct corollary based on previous discussions.

Corollary 1. *For any class of binary classifier $\mathcal{F}$ with $d = \mathrm{Ldim}(\mathcal{F})$, we have*

$$\mathcal{V}^{\mathrm{seq}}(\mathcal{F}, n) \le \mathcal{R}^{\mathrm{seq}}(\mathcal{F}) \le \sqrt{\frac{2d \ln\left(\frac{en}{d}\right)}{n}}.$$

So finite Littlestone dimension is sufficient for online learnability. It turns out that it is also *necessary* for online learnability, indicating that this sequence of upper bounding is tight.

Theorem 3. *If $\mathrm{Ldim}(\mathcal{F}) = \infty$, then for any algorithm and any n, there exists an environment such that the expected average regret of this algorithm is at least $1/2$.*

Proof. Let $\boldsymbol{x}$ be a tree of depth n shattered by $\mathcal{F}$ (which always exists since $\mathrm{Ldim}(\mathcal{F}) = \infty$), and $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables. At time t, the environment chooses $(\boldsymbol{x}_t(\boldsymbol{\epsilon}), \epsilon_t)$. In such an environment, no matter what the algorithm is (proper or improper), its expected total loss is always exactly $n/2$. On the other hand, by the definition of shattering, there is always an $f \in \mathcal{F}$ with perfect prediction on this dataset, which means that the expected regret in at least $n/2$. □

In fact, in HW2 you will prove an even stronger statement (with $d = \mathrm{Ldim}(\mathcal{F})$):

$$\mathcal{V}^{\mathrm{seq}}(\mathcal{F}, n) \ge \sqrt{\frac{1}{8}\min\left\{\frac{d}{n}, 1\right\}},$$

further showing that the upper bound we obtain is very tight. Closing the gap $\ln\left(\frac{en}{d}\right)$ between the upper and lower bounds remains open.

Summary. Combining all steps, we have shown for binary classification problems

$$\begin{aligned}\sqrt{\frac{1}{8}\min\left\{\frac{d}{n},1\right\}} &\le \mathcal{V}^{\text{seq}}(\mathcal{F},n)\\ &\le \sup_{\mathcal{P}\in\Delta(\mathcal{Z}^n)}\mathop{\mathbb{E}}_{z_{1:n}\sim\mathcal{P}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}_{z_t'\sim\mathcal{P}(\cdot|z_{1:t-1})}\left[\ell(f,z_t')\right]-\ell(f,z_t)\right)\right]\\ &\le 2\mathcal{R}^{\text{seq}}(\ell(\mathcal{F}))\le\mathcal{R}^{\text{seq}}(\mathcal{F})\le\sup_{\boldsymbol{x}}\sqrt{\frac{2\ln\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})}{n}}\le\sqrt{\frac{2d\ln\left(\frac{en}{d}\right)}{n}}\end{aligned}$$

where $d=\text{Ldim}(\mathcal{F})$.

2.2 Online learning is strictly harder

In Lecture 1, via the online-to-batch conversion we showed that online learning is at least as hard as statistical learning. Is it strictly harder? The example of the simple class defined by Equation (2) does not indicate that because the VC dimension and the Littlestone dimension coincide (both are 1). Instead, let's consider the threshold function class defined over $\mathcal{X}=\mathbb{R}$:

$$\mathcal{F}=\left\{f_\theta(x)=\begin{cases}+1 & \text{if } x\le\theta\\ -1 & \text{else}\end{cases}\;\middle|\;\theta\in\mathbb{R}\right\},\tag{3}$$

which has VC dimension exactly 1 as discussed in Lecture 2. It turns out that the Littlestone dimension of this seemingly simple class is infinity!

Proposition 1. *The Littlestone dimension of the threshold function class (Equation (3)) is ∞.*

Proof. To see this, consider an infinite $[0,1]$-valued tree $\boldsymbol{x}$ with root being $1/2$, and the left child and right child of a node with value a at level t being $a-\frac{1}{2^{t+1}}$ and $a+\frac{1}{2^{t+1}}$ respectively. (This is much easier to interpret if you draw a picture.)

Now for any n and any path/labeling $\boldsymbol{\epsilon}$, let θ_1 be the last node of this path $\boldsymbol{x}_n(\boldsymbol{\epsilon})$ and θ_2 be the last node on this path with label $-\epsilon_n$. Then the claim is that any value θ in between θ_1 and θ_2 satisfies: $f_\theta(\boldsymbol{x}_t(\boldsymbol{\epsilon}))=\epsilon_t$ for all $t=1,\dots,n$, and thus $\mathcal{F}$ shatters this tree. Indeed, note that the tree is constructed such that if $\epsilon_t=+1$, then every node in the path below level t has value larger than the node at level t. Similarly, if $\epsilon_t=-1$, then every node in the path below level t has value smaller than the node at level t. Therefore, suppose $\epsilon_n=+1$, then all the nodes with label $+1$ on the path must have a value smaller than θ_1, and all the nodes with label -1 on the path must have a value larger than θ_2, and thus f_θ predicts all the labels correctly if $\theta\in[\theta_1,\theta_2)$. The case for $\epsilon_n=-1$ is similar. □

Based on previous discussions, we conclude that this simple class is learnable in the statistical learning setting, but not learnable in the online setting. More generally, it is clear that the class of linear classifiers

$$\mathcal{F}=\left\{f_{\theta,b}(x)=\text{sign}\left(\langle x,\theta\rangle+b\right)\;|\;\theta\in\mathbb{R}^d,b\in\mathbb{R}\right\}$$

also has infinite Littlestone dimension (since it subsumes the threshold class), while having a finite VC dimension $d+1$. This illustrates that online learning is not just as hard as statistical learning, but is in fact *strictly* harder than statistical learning.

So if even learning linear classifiers is impossible, is online learning just too hard to be meaningful? The answer is yes in some sense for online classification, or put differently the 0-1 loss is too hard for online learning beyond finite classes. However, next time we will show that for online regression, many more possibilities open up.