CSCI 678: Theoretical Machine Learning
Lecture 3

Fall 2024, Instructor: Haipeng Luo

1 Infinite Class: Regression

In this lecture, we continue to focus on characterizing the learnability of statistical learning. Recall that in the last lecture, by a sequence of upper bounding we arrived at

$$\begin{aligned}
\mathcal{V}^{\text{iid}}(\mathcal{F}, n) &\le \sup_{\mathcal{P}} \left(\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left(L(f) - \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right) \right] \right) && \text{(using ERM)} \\
&\le 2 \sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\ell(\mathcal{F})) && \text{(symmetrization)} \\
&\le 2G \sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\mathcal{F}) && \text{(erasing the loss)}
\end{aligned}$$

where G is $1/2$ for a binary classification problem or the Lipschitz constant for the regression loss. For a finite class with bounded value in $[-C, C]$, we apply maximal inequality to show $\sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\mathcal{F}) \le C\sqrt{\frac{2 \ln |\mathcal{F}|}{n}}$. Based on this result, we further discussed that for a binary classification problem ($\mathcal{Y} = \{-1, +1\}$), since only the projection $\mathcal{F}|_{x_{1:n}} = \{(f(x_1), \cdots, f(x_n)) \mid f \in \mathcal{F}\} \subset \{-1, +1\}^n$ matters, one can reduce the infinite case to the finite case by introducing growth function and VC-dimension of a class. Specifically we proved with $d = \text{VCdim}(\mathcal{F})$,

$$\sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\mathcal{F}) \le \sqrt{\frac{2 \ln \Pi_{\mathcal{F}}(n)}{n}} \le \sqrt{\frac{2d \ln \left(\frac{en}{d}\right)}{n}}.$$

In this lecture, we turn our focus to regression problems with a real-valued function class. Without loss of generality, we assume that the output is normalized so that $\mathcal{Y} = [-1, +1]$ and $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$. It is clear that the key is still to understand the Rademacher complexity $\mathcal{R}^{\text{iid}}(\mathcal{F})$. However, since $\mathcal{F}$ is a real-valued class, the projection $\mathcal{F}|_{x_{1:n}} \subset [-1, +1]^n$ is generally also an infinite set and we cannot directly apply the finite case result.

A somewhat natural idea to fix this issue is to approximate the infinite class by a finite discretization. There are different possible ways to do this, and we discuss two below.

1.1 Covering functions

The first idea is to come up with a finite function class $\mathcal{H}$ so that for any $f \in \mathcal{F}$, there is a corresponding representative $h \in \mathcal{H}$ that is close to f. The closeness could be measured by for example $\sup_{x \in \mathcal{X}} |f(x) - h(x)|$. Based on this intuition, we define a *pointwise α-cover* of $\mathcal{F}$ as a finite class $\mathcal{H} \subset [-1, +1]^{\mathcal{X}}$ such that for any $f \in \mathcal{F}$, there exists $h \in \mathcal{H}$ such that $|f(x) - h(x)| \le \alpha$ for all $x \in \mathcal{X}$, and the pointwise α-covering number of $\mathcal{F}$ as

$$\mathcal{N}(\mathcal{F}, \alpha) = \min \{|\mathcal{H}| \mid \mathcal{H} \text{ is a pointwise } \alpha\text{-cover of } \mathcal{F}\}.$$

(or infinity if there is no such finite cover). Clearly, $\mathcal{N}(\mathcal{F}, \alpha)$ is non-increasing in α. With this definition, we can once again reduce the infinite case to the finite case and immediately derive the following result:

Theorem 1. *For any $\mathcal{F} \subset [-1,+1]^{\mathcal{X}}$, we have*

$$\mathcal{R}^{\text{iid}}(\mathcal{F}) \le \min_{\alpha \ge 0} \left(\alpha + \sqrt{\frac{2 \ln \mathcal{N}(\mathcal{F}, \alpha)}{n}} \right).$$

Proof. Fix any $\alpha \ge 0$. Let $\mathcal{H}$ be a pointwise α-cover of $\mathcal{F}$ with size $\mathcal{N}(\mathcal{F}, \alpha)$ and $h_f \in \mathcal{H}$ be the "representative" of $f \in \mathcal{F}$ such that $\sup_x |f(x) - h_f(x)| \le \alpha$. We then have

$$\begin{aligned}
\mathcal{R}^{\text{iid}}(\mathcal{F}) &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t)\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t (f(x_t) - h_f(x_t) + h_f(x_t))\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t (f(x_t) - h_f(x_t))\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t h_f(x_t)\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} |f(x_t) - h_f(x_t)|\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{H}} \sum_{t=1}^{n} \epsilon_t h(x_t)\right] \\
&\le \alpha + \mathcal{R}^{\text{iid}}(\mathcal{H}) \le \alpha + \sqrt{\frac{2 \ln \mathcal{N}(\mathcal{F}, \alpha)}{n}}. && \text{(Massart's lemma)}
\end{aligned}$$

Since this holds for any $\alpha \ge 0$, the theorem follows. □

Naturally, the bound exhibits some trade-off between the approximation scale α and the size of the cover. How large can the pointwise covering number be? Let's first consider a linear case, where $\mathcal{X} = B_q^d$ and $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_q^d \right\}$ for some $p \ge 1$ and $q \ge 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Here $B_p^d = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x\|_p \le 1 \right\}$ is the d-dimensional p-norm unit ball. The condition $\frac{1}{p} + \frac{1}{q} = 1$ makes $\|\cdot\|_p$ the dual norm of $\|\cdot\|_q$ (and vice versa), and this ensures that $|f_\theta(x)| = |\langle \theta, x \rangle| \le \|\theta\|_q \|x\|_p \le 1$ by Hölder inequality. This captures many common problems such as (regularized) linear regression. We first see how large the pointwise covering number is when $p = \infty$ (and thus $q = 1$).

Proposition 1. *For $\mathcal{X} = B_1^d$ and $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_\infty^d \right\}$, we have $\mathcal{N}(\mathcal{F}, \alpha) \le \left(\frac{1}{\alpha}\right)^d$ for any $0 \le \alpha \le 1$ and $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{d \ln\left(\frac{n}{d}\right)}{n}}\right)$ whenever $n \ge d$.*

Proof. Note that B_∞^d is simply a d-dimensional hypercube with edge length 2. Fix any $0 \le \alpha \le 1$. We discretize this hypercube "evenly" into $\left(\frac{1}{\alpha}\right)^d$ disjoint small hypercubes with edge length 2α,[1] and define $\mathcal{H} \subset \mathcal{F}$ as a set of linear functions parametrized by the centers of these small hypercubes. Clearly, $\mathcal{H}$ is a pointwise α-cover of $\mathcal{F}$, since for any $f_\theta \in \mathcal{F}$, if we let θ' be the center of the small hypercube that θ lies in and $h_{\theta'} \in \mathcal{H}$ be the corresponding linear function, then for any $x \in B_1^d$ we have $|f_\theta(x) - h_{\theta'}(x)| = |\langle \theta - \theta', x \rangle| \le \|\theta - \theta'\|_\infty \|x\|_1 \le \alpha$. This concludes the first statement. The second statement is by applying Theorem 1 and setting $\alpha = \sqrt{d/n}$. □

This implies that the linear class above is learnable (via ERM with rate roughly $\sqrt{d/n}$). For a general value of p, it is easy to see that $B_p^d \subset B_\infty^d$. So if we use the same class $\mathcal{H}$ constructed in the proof of Proposition 1 as a pointwise cover, we can show that for any f_θ and its representative $h_{\theta'}$, one has for any $x \in B_q^d$

$$|f_\theta(x) - h_{\theta'}(x)| = |\langle \theta - \theta', x \rangle| \le \|\theta - \theta'\|_p \|x\|_q \le d^{\frac{1}{p}} \|\theta - \theta'\|_\infty \le d^{\frac{1}{p}} \alpha,$$

which means that $\mathcal{H}$ is a pointwise $d^{\frac{1}{p}}\alpha$-cover and consequentially the pointwise covering number is bounded as $\mathcal{N}(\mathcal{F}, \alpha) \le \left(\frac{d^{\frac{1}{p}}}{\alpha}\right)^d$. So the linear class is learnable for any value of p.

[1] Technically, $\left(\frac{1}{\alpha}\right)^d$ should be $\lceil \frac{1}{\alpha} \rceil^d$. We ignore this subtlety since it makes no real difference.

However, when dealing with the p-norm ball B_p^d, intuitively we should also discretize it into small p-norm balls instead of small hypercubes, and this might lead to a smaller cover. This is indeed true as shown in the next proposition, but explicitly constructing such a cover seems rather difficult. Fortunately, in the proof we show that sometimes it is possible to give a bound on the covering number without explicitly constructing the cover.

Proposition 2. *If $\mathcal{X} = B_q^d$ and $\mathcal{F} = \left\{f_\theta(x) = \langle \theta, x\rangle \mid \theta \in B_p^d\right\}$ for some $p \geq 1$ and $q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have $\mathcal{N}(\mathcal{F}, \alpha) \leq \left(\frac{2}{\alpha} + 1\right)^d$ for any $0 \leq \alpha \leq 1$ and $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{d \ln\left(\frac{n}{d}\right)}{n}}\right)$ whenever $n \geq d$.*

Proof. Fix any $0 \leq \alpha \leq 1$. Let rB_p^d be a p-norm ball with radius r for some $r \geq 0$. The key idea is to *pack* the ball B_p^d with as many small balls $\frac{\alpha}{2}B_p^d$ as possible. Formally, let $S \subset B_p^d$ be the largest subset such that for any two points $\theta, \theta' \in S$, we have $\|\theta - \theta'\|_p > \alpha$ (this is call an *α-packing* of B_p^d). We first claim that the corresponding function class $\mathcal{H} = \{h_\theta(x) = \langle \theta, x\rangle \mid \theta \in S\}$ is a pointwise α-cover of $\mathcal{F}$. Indeed, for any $\theta \in B_p^d$, there must exist $\theta' \in S$ such that $\|\theta - \theta'\|_p \leq \alpha$, since otherwise θ can be added to S and it is still an α-packing, a contradiction to the definition of S. It is then clear that $|f_\theta(x) - h_{\theta'}(x)| \leq \alpha$ for any $x \in B_q^d$.

It remains to prove $|S| \leq \left(\frac{2}{\alpha} + 1\right)^d$. To show this, imagine that for each point in S, we put a p-norm ball with radius $\frac{\alpha}{2}$ centered at this point. By the definition of S, all these balls are disjoint. On the other hand, all these balls are contained in the larger ball $(1 + \frac{\alpha}{2})B_p^d$. Therefore, we must have that the sum of the volumes of all these small balls $\frac{\alpha}{2}B_p^d$ is bounded by the volume of the larger ball $(1 + \frac{\alpha}{2})B_p^d$: $|S|\text{Vol}(\frac{\alpha}{2}B_p^d) \leq \text{Vol}((1 + \frac{\alpha}{2})B_p^d)$. Using the fact $\text{Vol}(rB_p^d) = r^d\text{Vol}(B_p^d)$ and rearranging then proves $|S| \leq \left(\frac{2}{\alpha} + 1\right)^d$. The upper bound on $\mathcal{R}^{\text{iid}}(\mathcal{F})$ is again obtained by applying Theorem 1 and setting $\alpha = \sqrt{d/n}$. □

In HW1, you will also prove that this covering number of order $\mathcal{O}(\frac{1}{\alpha^d})$ is tight for the linear class, using a similar volumetric argument. Next, we consider a *nonparametric* example where $\mathcal{X} = \mathbb{R}$ and $\mathcal{F}$ is the set of *all non-decreasing functions*. This function class is commonly used in the so-called isotonic regression problems, where it is very natural to assume that the output is monotonic in the input (for example, predicting the height of children as a function of age). In this case, $\mathcal{F}$ seems to be a very expressive class. Indeed, it has an infinite pointwise covering number, as shown below.

Proposition 3. *If $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [-1, +1]$, and $\mathcal{F} \in \mathcal{Y}^\mathcal{X}$ is the set of all non-decreasing functions, then $\mathcal{N}(\mathcal{F}, \alpha) = \infty$ for any $\alpha < 1$.*

Proof. Consider an infinite subset of $\mathcal{F}$ defined as $\{f_m(x) = \text{sign}(x - m) \mid m \text{ is an integer}\}$. It is impossible to pointwise cover any two different functions f_m and $f_{m'}$ from this set with the same function h, since $|f_m(\frac{m+m'}{2}) - f_{m'}(\frac{m+m'}{2})| = 2$ and thus $h(\frac{m+m'}{2})$ cannot be simultaneously α-close to both $f_m(\frac{m+m'}{2})$ and $f_{m'}(\frac{m+m'}{2})$. This implies that there is no finite pointwise cover for $\mathcal{F}$. □

Does this imply that this function class is not learnable? The answer is no as we show in the next section. Importantly, this implies that pointwise covering is in fact not the right, or at least not the tight, complexity measure.

1.2 Covering projections

Recall that just as in classification, symmetrization allows us to only care about the projection $\mathcal{F}|_{x_{1:n}}$, instead of the entire function class $\mathcal{F}$. This motivates us to approximately discretizes the n-dimensional space $\mathcal{F}|_{x_{1:n}}$ instead of $\mathcal{F}$. Formally, we say that $V \subset [-1, +1]^n$ is an α-cover of $\mathcal{F}|_{x_{1:n}}$ with respect to ℓ_∞ norm if for any $f \in \mathcal{F}|_{x_{1:n}}$, there exists $v \in V$ such that $\|f - v\|_\infty \leq \alpha$. Note that here we slightly abuse the notation by using f as an n-dimensional vector (while previously it was also used as a function in $\mathcal{F}$). The (ℓ_∞ norm) α-covering number $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha)$ is defined as the size of the smallest ℓ_∞ norm α-cover.

In fact, a more careful inspection of the proof of Theorem 1 reveals that ℓ_∞ cover is not necessarily needed. To this end, for any $p > 0$, we say that $V \subset [-1,+1]^n$ is an α-cover of $\mathcal{F}|_{x_{1:n}}$ with respect to ℓ_p norm if for any $f \in \mathcal{F}|_{x_{1:n}}$, there exists $v \in V$ such that $\|f - v\|_p \le n^{\frac{1}{p}}\alpha$, or equivalently

$$\left(\frac{1}{n}\sum_{t=1}^{n}|f_t - v_t|^p\right)^{\frac{1}{p}} \le \alpha.$$

Similarly, the corresponding α-covering number $\mathcal{N}_p(\mathcal{F}|_{x_{1:n}}, \alpha)$ is defined as the size of the smallest ℓ_p norm α-cover.

Note that there is a somewhat "strange" (but in fact conventional) normalization going on in this definition. This normalization ensures that $\left(\frac{1}{n}\sum_{t=1}^{n}|f_t - v_t|^p\right)^{\frac{1}{p}}$ is an increasing function in p (you can prove this via Hölder inequality), and therefore

$$\mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha) \le \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha) \le \cdots \le \mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha).$$

An argument similar to the proof of Theorem 1 shows the following.

Theorem 2. *For any $\mathcal{F} \subset [-1,+1]^{\mathcal{X}}$ and inputs $x_{1:n}$, we have*

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le \min_{\alpha \ge 0}\left(\alpha + \sqrt{\frac{2\ln \mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha)}{n}}\right).$$

Proof. Fix any $\alpha \ge 0$. Let V be an α-cover of $\mathcal{F}|_{x_{1:n}}$ with size $\mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha)$ and $v_f \in V$ be the "representative" of $f \in \mathcal{F}|_{x_{1:n}}$ such that $\|f - v_f\|_1 \le n\alpha$. We then have with $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$,

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) &= \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \langle \epsilon, f\rangle\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \langle \epsilon, f - v_f + v_f\rangle\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \langle \epsilon, f - v_f\rangle\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \langle \epsilon, v_f\rangle\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \|f - v_f\|_1\right] + \frac{1}{n}\mathbb{E}\left[\sup_{v \in V} \langle \epsilon, v\rangle\right] \\
&\le \alpha + \sqrt{\frac{2\ln \mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha)}{n}}. \qquad \text{(Massart's lemma)}
\end{aligned}$$

Since this holds for any $\alpha \ge 0$, the theorem follows. □

Now let's see how covering projections is better than covering functions. First, since $\mathcal{F}|_{x_{1:n}}$ lies in $[-1,+1]^n$, it is trivial to see that by discretizing $[-1,+1]^n$ into small hypercubes, similarly to what we did in the previous section, one can show that $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha) \le \left(\frac{1}{\alpha}\right)^n$ always holds. This bound is useless though since it leads to a constant upper bound on the Rademacher complexity if one plugs this into the bound of Theorem 2.

Second, note that if $\mathcal{H}$ is a pointwise α-cover of $\mathcal{F}$, then by definition $\mathcal{H}|_{x_{1:n}}$ is also a α-cover of $\mathcal{F}|_{x_{1:n}}$ with respect to ℓ_∞ norm, which implies $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha) \le \mathcal{N}(\mathcal{F}, \alpha)$ (that is, covering projections is never worse than covering functions). Therefore, for the linear class $\mathcal{F} = \left\{f_\theta(x) = \langle \theta, x\rangle \mid \theta \in B_p^d\right\}$ discussed earlier, one also has $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha) \le \left(\frac{2}{\alpha} + 1\right)^d$. In fact, it is not hard to see that more generally, as long as $\mathcal{F}|_{x_{1:n}}$ lies in some d-dimensional subspace of $[-1,+1]^n$, then its covering number is roughly of order $\mathcal{O}((\frac{1}{\alpha})^d)$ (see HW1).

Finally, we come back to the non-decreasing function class and argue that while $\mathcal{N}(\mathcal{F}, \alpha) = \infty$, $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha)$ is finite, which means covering projections is strictly better than covering functions.

Proposition 4. *If $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [-1,+1]$, and $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$ is the set of all non-decreasing functions, then $\mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \alpha) \le (n+1)^{\frac{1}{\alpha}}$ for any $\alpha < 1$, and thus*

$$\mathcal{R}^{\text{iid}}(\mathcal{F}) \le \min_{0 \le \alpha \le 1}\left(\alpha + \sqrt{\frac{2\ln(n+1)}{\alpha n}}\right) = \mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{3}}\right).$$

Proof. Without loss of generality we assume $x_1 \le \cdots \le x_n$. Let $S \subset [-1,+1]$ be a finite discretization at scale 2α such that $|S| \le 1/\alpha$ and for any $y \in [-1,+1]$, there exists $y' \in S$ such that $|y - y'| \le \alpha$. Let $V = \{v \in S^n \mid v_1 \le \cdots \le v_n\}$. Clearly, by construction V is an α-cover of $\mathcal{F}|_{x_{1:n}}$ with respect to ℓ_∞ norm. It remains to calculate the size of V. It is not hard to see that $|V|$ is exactly the number of solutions of the equation $\sum_{i=1}^{|S|} m_i = n$ for non-negative integers $m_1, \dots, m_{|S|}$, where m_i represents the number of appearances of the ith-smallest element of S. The exact number of solutions is $\binom{n+|S|-1}{|S|-1}$, but a rough estimate $(n+1)^{|S|} \le (n+1)^{\frac{1}{\alpha}}$ can be obtained by simply realizing that each m_i can only take $n+1$ possible values. The bound on the Rademacher complexity is by a direct application of Theorem 2 and picking the optimal value of α. □

This shows that while the class of all non-decreasing functions is seemingly very expressive, it is in fact still learnable via ERM. This serves as another example to showcase the importance of the symmetrization trick, which allows us to focus only on the projections but not the functions.

We finally remark that no matter what kind of covers we are using, it only happens in the analysis but not the algorithm — the algorithm is always just ERM, which could be very efficient even for problems like isotonic regression.

2 Dudley Entropy Integral

One might notice that the rate of convergence shown in Proposition 6 is roughly $1/n^{\frac{1}{3}}$, which is slower than the typical rate $1/\sqrt{n}$ we have seen for all other examples. Does that really imply that learning non-decreasing functions requires more samples, or is our bound loose? It turns out that the latter is true, and to improve the bound, we need to apply a tighter analysis using the so-called *Dudley entropy integral*.

Theorem 3. *For any $\mathcal{F} \subset [-1,+1]^{\mathcal{X}}$ and inputs $x_{1:n}$, we have*

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le \min_{0 \le \alpha \le 1} \left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta \right).$$

The proof is deferred to the next section. The upper bound in the theorem above is called the Dudley entropy integral of class $\mathcal{F}$ (log covering number is often called the metric entropy; hence the name). It is in terms of the ℓ_2 covering number (the reason will be clear in the proof), and it looks at the covering number at different scales simultaneously. Ignoring constants and the difference between $\mathcal{N}_1$ and $\mathcal{N}_2$, this is never worse than the bound given by Theorem 2 since $\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)}$ is decreasing in δ and thus $\int_\alpha^1 \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta \le (1-\alpha)\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha)} \le \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha)}$. It could be strictly better though as shown in the following two examples.

Proposition 5. *If $\mathcal{X} = B_q^d$ and $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_p^d \right\}$ for some $p \ge 1$ and $q \ge 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}\left(\sqrt{d/n}\right)$.*

Proof. We use the bound $\mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta) \le \mathcal{N}(\mathcal{F}, \delta) \le \left(\frac{2}{\alpha} + 1\right)^d \le \left(\frac{3}{\alpha}\right)^d$, and thus $\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} \le \sqrt{d \ln\left(\frac{3}{\delta}\right)} \le 2\sqrt{d} \ln\left(\frac{1}{\delta}\right)$ for $\delta \le 1/3$. Therefore we have

$$\begin{aligned}
\int_\alpha^1 \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta &= \int_\alpha^{1/3} \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta + \int_{1/3}^1 \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta \\
&\le 2\sqrt{d} \int_\alpha^{1/3} \ln\left(\frac{1}{\delta}\right) d\delta + \mathcal{O}(\sqrt{d}) \\
&= 2\sqrt{d} \left(\delta - \delta \ln \delta\right) \Big|_\alpha^{1/3} + \mathcal{O}(\sqrt{d}).
\end{aligned}$$

Setting $\alpha = 0$ finishes the proof. □

So using the Dudley entropy integral allows us the remove the extra $\ln n$ term (in Proposition 2) for the Rademacher complexity of linear functions. The improvement in the next example will be even more significant.

Proposition 6. *If $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = [-1, +1]$, and $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$ is the set of all non-decreasing functions, then $\mathcal{R}^{\text{iid}}(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{\ln n}{n}}\right)$.*

Proof. Again we directly plug in the bound $\mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta) \le \mathcal{N}_\infty(\mathcal{F}|_{x_{1:n}}, \delta) \le (n+1)^{\frac{1}{\delta}}$ and calculate the Dudley entropy integral:

$$\int_\alpha^1 \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta \le \sqrt{\ln(n+1)} \int_\alpha^1 \frac{1}{\sqrt{\delta}} d\delta = 2\sqrt{\ln(n+1)}(1 - \sqrt{\alpha}) \le 2\sqrt{\ln(n+1)}.$$

Setting $\alpha = 0$ finishes the proof. □

This shows that the rate for learning non-decreasing functions is again roughly $1/\sqrt{n}$ instead of $1/n^{\frac{1}{3}}$, demonstrating the power of Dudley entropy integral.

2.1 Chaining Technique

Proof of Theorem 3. The proof relies on an important *chaining* technique that looks at different scales of covering simultaneously. Specifically, for $j = 1, 2, \ldots, M$ (for some M to be specified later), let $\alpha_j = 2^{-j}$ and V_j be an (ℓ_2 norm) α_j-cover of $\mathcal{F}|_{x_{1:n}}$ with size $\mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha_j)$. Then for each $f \in \mathcal{F}|_{x_{1:n}}$, we can associate it with a chain of representatives $v_f^j \in V_j$ for $j = 1, 2, \ldots, M$ such that $\left\|f - v_f^j\right\|_2 \le \sqrt{n}\alpha_j$. Additionally, we let v_f^0 be the all-zero vector for notational convenience. We then have with $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$,

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) &= \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \langle \epsilon, f \rangle\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \left\langle \epsilon, f - v_f^M \right\rangle + \sum_{j=1}^M \left\langle \epsilon, v_f^j - v_f^{j-1} \right\rangle\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \left\langle \epsilon, f - v_f^M \right\rangle\right] + \frac{1}{n}\sum_{j=1}^M \mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \left\langle \epsilon, v_f^j - v_f^{j-1} \right\rangle\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}|_{x_{1:n}}} \left\|f - v_f^M\right\|_1\right] + \frac{1}{n}\sum_{j=1}^M \mathbb{E}\left[\sup_{(v,v') \in S_j} \langle \epsilon, v - v' \rangle\right],
\end{aligned}$$

where

$$S_j = \left\{ (v, v') \in V_j \times V_{j-1} \;\middle|\; \exists\, f \in \mathcal{F}|_{x_{1:n}} \text{ s.t. } v \text{ and } v' \text{ are both representatives of } f \right\}.$$

The first term in the last bound is bounded by α_M since $\left\|f - v_f^M\right\|_1 \le \sqrt{n}\left\|f - v_f^M\right\|_2 = n\alpha_M$ by Cauchy-Schwarz inequality. For the second term, we apply Massart's lemma again:

$$\mathbb{E}\left[\sup_{(v,v') \in S_j} \langle \epsilon, v - v' \rangle\right] \le \sigma\sqrt{2\ln(|V_j||V_{j-1}|)} \le 2\sigma\sqrt{\ln|V_j|} = 2\sigma\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha_j)}$$

where $\sigma = \sup_{(v,v') \in S_j} \|v - v'\|_2$. Since for any pair $(v, v') \in S_j$, there exists f such that $\|v - f\|_2 \le \sqrt{n}\alpha_j$ and $\|v' - f\|_2 \le \sqrt{n}\alpha_{j-1}$, one has

$$\|v - v'\|_2 \le \|v - f\|_2 + \|v' - f\|_2 \le \sqrt{n}(\alpha_j + \alpha_{j-1}) = 3\sqrt{n}\alpha_j.$$

This shows $\sigma \le 3\sqrt{n}\alpha_j$ and thus

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) &\le \alpha_M + \frac{6}{\sqrt{n}}\sum_{j=1}^M \alpha_j \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha_j)} \\
&\le \alpha_M + \frac{12}{\sqrt{n}}\sum_{j=1}^M (\alpha_j - \alpha_{j+1})\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha_j)} \\
&\le \alpha_M + \frac{12}{\sqrt{n}}\int_{\alpha_{M+1}}^{\alpha_1} \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta,
\end{aligned}$$

where the last step uses the fact that $\sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)}$ is a non-increasing function in δ. Finally, for any $0 \le \alpha \le 1$, let M be such that $2^{-(M+2)} \le \alpha \le 2^{-(M+1)}$, then we have $\alpha_M \le 4\alpha$ and $\alpha \le \alpha_{M+1}$, and thus

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta,$$

finishing the proof. □

Note that the key reason that ℓ_2 norm covers are used is because the definition of σ is in terms of ℓ_2 norm (which is inherited from the maximal inequality).

Summary To recap, we have shown three different upper bounds in terms of covering numbers for the Rademacher complexity of a real-valued class, along with two running examples to show how good each bound is; see the table below for a summary.

Table 1: Summary of Rademacher complexity upper bounds using covering numbers

	Upper bounds	**Examples**	
		linear functions	non-decreasing functions
$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le$	$\min_{\alpha \ge 0} \left(\alpha + \sqrt{\frac{2 \ln \mathcal{N}(\mathcal{F}, \alpha)}{n}} \right)$	$\mathcal{O}\left(\sqrt{\frac{d \ln\left(\frac{n}{d}\right)}{n}} \right)$	∞
	$\min_{\alpha \ge 0} \left(\alpha + \sqrt{\frac{2 \ln \mathcal{N}_1(\mathcal{F}|_{x_{1:n}}, \alpha)}{n}} \right)$	$\mathcal{O}\left(\sqrt{\frac{d \ln\left(\frac{n}{d}\right)}{n}} \right)$	$\mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{\frac{1}{3}} \right)$
	$\min_{0 \le \alpha \le 1} \left(4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \delta)} d\delta \right)$	$\mathcal{O}\left(\sqrt{\frac{d}{n}} \right)$	$\mathcal{O}\left(\sqrt{\frac{\ln n}{n}} \right)$

3 Combinatorial Parameters: Pseudo-Dimension

Note that the role of covering number is very similar to the role of growth function for classification problems. For the latter, 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 $f(x) \le y$ (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 drawing a picture for the case $\mathcal{X} = \mathbb{R}$ to help understand this.)

Take the linear class as an example again: $\mathcal{X} = B_q^d$, $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_p^d \right\}$ for some $p \ge 1$ and $q \ge 1$ 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 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 those in Table 1.

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 already proved), 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.

How do we fix this? Is there a better combinatorial parameter whose finiteness is necessary for learning? We will answer these questions in the next lecture.