CSCI 678: Theoretical Machine Learning
Lecture 2

Fall 2024, Instructor: Haipeng Luo

1 Uniform Convergence and Rademacher Complexity

In this lecture, we focus on studying the value $\mathcal{V}^{\text{iid}}(\mathcal{F}, n)$, which, as discussed last time, completely characterizes the learnability of class $\mathcal{F}$ in the batch/statistical learning setting. We will perform a sequence of upper bounding on this value to reach a much more manageable form, and in the end argue that these upper bounds are very tight.

Recall that the value is defined as

$$\mathcal{V}^{\text{iid}}(\mathcal{F}, n) = \inf_{\pi} \sup_{\mathcal{P}} \left(\mathbb{E}\left[L(\widehat{y})\right] - \inf_{f \in \mathcal{F}} L(f) \right),$$

where π ranges over all (distributions of) mappings from n training samples to a final predictor $\widehat{y} \in \mathcal{D}$, and $\mathcal{P}$ ranges over all data-generating distributions on $\mathcal{Z}$. As the first step of relaxing this value, we consider a very simple algorithm: output the *Empirical Risk Minimizer* (ERM):

$$\widehat{y}_{\text{ERM}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t).$$

Here, the empirical risk simply refers to the average loss over the training set, and for simplicity we assume that at least one such minimizer exists (which is basically without loss of generality). Clearly, we now have

$$\mathcal{V}^{\text{iid}}(\mathcal{F}, n) \leq \sup_{\mathcal{P}} \left(\mathbb{E}\left[L(\widehat{y}_{\text{ERM}})\right] - \inf_{f \in \mathcal{F}} L(f) \right). \tag{1}$$

Before further discussion, we make the following two remarks on ERM. First, finding an ERM is a well-defined optimization problem, and there are many heavily-studied optimization algorithms for this problem. As discussed last time, this optimization aspect is out of the scope of this course. In general, finding an ERM could even be an NP-hard problem. However, here we only focus on whether the problem is statistically (as opposed to computationally) learnable.

Second, one might wonder why we should focus on this somewhat "naive" algorithm, or more specifically, why is there no "regularization" as we know that minimizing training loss alone might lead to overfitting in practice. The answer is that for many problems, a regularized ERM, that is, $\operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) + \lambda \Psi(f)$ for some regularizer Ψ and constant λ, is equivalent to the ERM over a smaller class: $\operatorname{argmin}_{f \in \mathcal{F}, \Psi(f) \leq c} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t)$ for some other constant c. So regularization is really just a way to implicitly learn over a restricted class that hopefully is statistically easier to learn.

1.1 Empirical process and uniform convergence

Next, we further simplify Equation (1). Let $f^\star \in \operatorname{argmin}_{f\in\mathcal{F}} L(f)$.[1] We have

$$\begin{aligned}
\mathcal{V}^{\text{iid}}(\mathcal{F}, n) &\le \sup_{\mathcal{P}} \left(\mathbb{E}\left[L(\widehat{y}_{\text{ERM}})\right] - L(f^\star)\right) \\
&= \sup_{\mathcal{P}} \mathbb{E}\left[L(\widehat{y}_{\text{ERM}}) - \frac{1}{n}\sum_{t=1}^n \ell(f^\star, z_t)\right] && \text{(by definition of } L\text{)} \\
&\le \sup_{\mathcal{P}} \mathbb{E}\left[L(\widehat{y}_{\text{ERM}}) - \frac{1}{n}\sum_{t=1}^n \ell(\widehat{y}_{\text{ERM}}, z_t)\right] && \text{(by definition of } \widehat{y}_{\text{ERM}}\text{)} \\
&\le \sup_{\mathcal{P}} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)\right)\right]. && (2)
\end{aligned}$$

Here, the collection of random variables $L(f) - \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)$ indexed by f is called an *empirical process*. Each of these random variables is nothing but the difference between the expected value of f on a random input drawn from $\mathcal{P}$ and its empirical average value on a set of i.i.d. inputs drawn from the same distribution. Clearly, these random variables are all zero-mean and each should be small when n is large, by the law of large numbers. However, in order to claim that Equation (2) is small, in some sense we have to argue that these random variables are all small simultaneously, and whether this is true will depend on the class $\mathcal{F}$. We say that $\mathcal{F}$ satisfies *uniform convergence* if

$$\lim_{n\to\infty} \sup_{\mathcal{P}} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)\right)\right] = 0,$$

that is, for any data-generating distribution, the expected supremum of the empirical process is arbitrarily small as long as n is large enough. By Equation (2), it is clear that if uniform convergence holds for $\mathcal{F}$, then $\mathcal{F}$ is learnable.

1.2 Symmetrization and Rademacher Complexity

For a given class $\mathcal{F}$, how do we know if the expected supremum of the empirical process is small or not? To answer this question, we will further relax this quantity via an important technique called *symmetrization*, and arrive at something called Rademacher complexity. To this end, we first define a *Rademacher random variable* ϵ as a random variable that takes on values -1 and $+1$ with equal probability. For a class of functions $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$, and a sequence of arbitrary inputs $z_1, \ldots, z_n$, define the *conditional Rademacher complexity* of $\mathcal{H}$ on these inputs as

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{H}; z_{1:n}) = \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^n \epsilon_t h(z_t)\right],$$

where $\epsilon_{1:n}$ are n i.i.d. Rademacher random variables. The (unconditional) *Rademacher complexity* of $\mathcal{H}$ with respect to a distribution $\mathcal{P}$ supported on $\mathcal{Z}$ is defined as

$$\mathcal{R}^{\text{iid}}(\mathcal{H}) = \mathbb{E}_{z_{1:n}}\left[\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{H}; z_{1:n})\right] = \frac{1}{n}\mathbb{E}_{z_{1:n},\epsilon_{1:n}}\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^n \epsilon_t h(z_t)\right]$$

where $z_{1:n}$ are n i.i.d. samples of $\mathcal{P}$. At a high level, the Rademacher complexity of a class measures how well it can fit random signs, because the correlation $\epsilon_t h(z_t)$ is large when $h(z_t)$ is of the same sign as ϵ_t. Therefore, the larger the Rademacher complexity, the more expressive the class is.

The connection between the Rademacher complexity and the expected supremum of an empirical process is summarized in the following theorem.

Theorem 1. *For any data-generating distribution $\mathcal{P}$ and any class $\mathcal{F}$, we have*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)\right)\right] \le 2\mathcal{R}^{\text{iid}}(\ell(\mathcal{F})),$$

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

[1] For simplicity, we will ignore the issue that argmin might not exist (which can be handled easily).

Proof. By definition, we write $L(f)$ as $\mathbb{E}_{z'_1,\ldots,z'_n\sim\mathcal{P}}\left[\frac{1}{n}\sum_{t=1}^n \ell(f,z'_t)\right]$ and arrive at

$$\mathbb{E}_{z_{1:n}}\left[\sup_{f\in\mathcal{F}}\left(L(f)-\frac{1}{n}\sum_{t=1}^n \ell(f,z_t)\right)\right] = \frac{1}{n}\mathbb{E}_{z_{1:n}}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{z'_{1:n}}\left[\sum_{t=1}^n\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right]\right]$$

Pulling the expectation $\mathbb{E}_{z'_{1:n}}$ out of the $\sup$ leads to the following upper bound (that is symmetric):

$$\frac{1}{n}\mathbb{E}_{z_{1:n}}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{z'_{1:n}}\left[\sum_{t=1}^n\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right]\right] \le \frac{1}{n}\mathbb{E}_{z_{1:n},z'_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right].$$

Next we claim the following equality:

$$\mathbb{E}_{z_{1:n},z'_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right] = \mathbb{E}_{z_{1:n},z'_{1:n},\epsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right].$$

This is true because for each possible value of the sequence $\epsilon_{1:n}\in\{-1,+1\}^n$, the difference between $\sum_{t=1}^n\left(\ell(f,z'_t)-\ell(f,z_t)\right)$ and $\sum_{t=1}^n\epsilon_t\left(\ell(f,z'_t)-\ell(f,z_t)\right)$ is simply that we switch the two examples z_t and z'_t whenever $\epsilon_t=-1$, and this makes no difference in expectation since z_t and z'_t follow the same distribution. Splitting the "sup" into two parts then further leads to the following upper bound:

$$\begin{aligned}
&\mathbb{E}_{z_{1:n},z'_{1:n},\epsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\left(\ell(f,z'_t)-\ell(f,z_t)\right)\right]\\
&\le \mathbb{E}_{z_{1:n},z'_{1:n},\epsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\ell(f,z'_t)+\sup_{f\in\mathcal{F}}\sum_{t=1}^n-\epsilon_t\ell(f,z_t)\right]\\
&= \mathbb{E}_{z'_{1:n},\epsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\ell(f,z'_t)\right]+\mathbb{E}_{z_{1:n},\epsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n-\epsilon_t\ell(f,z_t)\right].
\end{aligned}$$

Finally, noting that ϵ_t and $-\epsilon_t$ have the same distribution so the two terms above are both exactly $n\mathcal{R}^{\text{iid}}(\ell(\mathcal{F}))$ finishes the proof. □

1.3 Erasing the loss for supervised learning

For many problems, especially those for supervised learning, it turns out that when analyzing the Rademacher complexity of the class $\ell(\mathcal{F})$, the part about the loss function is in fact not that important and can usually be removed. Specifically, consider a supervised learning problem with $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and $\mathcal{F}\subset\mathcal{Y}^{\mathcal{X}}$. In the following two cases, we can easily relate $\mathcal{R}^{\text{iid}}(\ell(\mathcal{F}))$ and $\mathcal{R}^{\text{iid}}(\mathcal{F})$.

Lemma 1. *For a binary classification problem with $\mathcal{Y}=\{-1,+1\}$ and 0-1 loss, one has $\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F});z_{1:n})=\frac{1}{2}\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F};x_{1:n})$ for any sequence $z_{1:n}$, and thus $\mathcal{R}^{\text{iid}}(\ell(\mathcal{F}))=\frac{1}{2}\mathcal{R}^{\text{iid}}(\mathcal{F})$.*

Proof. By definition, we have

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F});z_{1:n}) &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\mathbb{I}\{f(x_t)\neq y_t\}\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n\epsilon_t\frac{1-y_tf(x_t)}{2}\right]\\
&= \frac{1}{2n}\mathbb{E}\left[\sum_{t=1}^n\epsilon_t\right]+\frac{1}{2n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n-\epsilon_ty_tf(x_t)\right] = \frac{1}{2}\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F};x_{1:n}),
\end{aligned}$$

where the last step uses the fact that Rademacher variables are zero-mean and that for any labels $y_1,\ldots,y_n$, $-\epsilon_1y_1,\ldots,-\epsilon_ny_n$ are again i.i.d. Rademacher random variables. □

Lemma 2 (Contraction lemma)**.** *For a regression problem with $\mathcal{Y}\subset\mathbb{R}$ and loss $\ell(f,(x,y))=\ell'(f(x),y)$ for some loss $\ell'(y',y)$ that is G-Lipschitz in the first parameter (that is, $|\ell'(y_1,y)-\ell'(y_2,y)|\le G|y_1-y_2|$ for any y_1,y_2 and y), one has $\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F});z_{1:n})\le G\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F};x_{1:n})$ for any sequence $z_{1:n}$, and thus $\mathcal{R}^{\text{iid}}(\ell(\mathcal{F}))\le G\mathcal{R}^{\text{iid}}(\mathcal{F})$.*

Proof. By definition, we have

$$\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F}); z_{1:n}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t\ell'(f(x_t),y_t)\right]$$

$$= \frac{1}{2n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t\ell'(f(x_t),y_t) + \ell'(f(x_n),y_n))\right] + \frac{1}{2n}\mathbb{E}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t\ell'(g(x_t),y_t) - \ell'(g(x_n),y_n))\right]$$

$$= \frac{1}{2n}\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t(\ell'(f(x_t),y_t) + \epsilon_t\ell'(g(x_t),y_t) + \ell'(f(x_n),y_n) - \ell'(g(x_n),y_n))\right]$$

$$\leq \frac{1}{2n}\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t(\ell'(f(x_t),y_t) + \epsilon_t\ell'(g(x_t),y_t) + G|f(x_n) - g(x_n)|)\right]$$

where the last step uses the G-Lipschitzness of ℓ'. By symmetry, removing the the absolute value in the last expression in fact makes no difference. Splitting the "sup" again we thus arrive at:

$$\widehat{\mathcal{R}}^{\text{iid}}(\ell(\mathcal{F}); z_{1:n}) \leq \frac{1}{2n}\mathbb{E}\left[\sup_{f,g\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t(\ell'(f(x_t),y_t) + \epsilon_t\ell'(g(x_t),y_t) + Gf(x_n) - Gg(x_n))\right]$$

$$= \frac{1}{2n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t\ell'(f(x_t),y_t) + Gf(x_n))\right] + \frac{1}{2n}\mathbb{E}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{n-1}(\epsilon_t\ell'(g(x_t),y_t) - Gg(x_n))\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}\epsilon_t\ell'(f(x_t),y_t) + \epsilon_n Gf(x_n)\right].$$

Keep doing this for $t = n-1, \ldots, 1$ finishes the proof. □

Note that Lipschitzness is usually satisfied for common problems. Take square loss $\ell'(y', y) = (y'-y)^2$ as an example. If $\mathcal{Y} = [-1, +1]$ then it is clear that the loss is 4-Lipschitz.

2 Finite Class

Let's make a quick summary at this point. By a sequence of upper bounding, we relax the value of a statistical learning problem to the Rademacher complexity of the class:

$$\mathcal{V}^{\text{iid}}(\mathcal{F}, n) \leq \sup_{\mathcal{P}}\left(\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^{n}\ell(f, z_t)\right)\right]\right) \leq 2\sup_{\mathcal{P}}\mathcal{R}^{\text{iid}}(\ell(\mathcal{F})) \leq 2G\sup_{\mathcal{P}}\mathcal{R}^{\text{iid}}(\mathcal{F})$$

where G is $1/2$ for a binary classification problem or the Lipschitz constant for the regression loss. It is now clear that understanding the Rademacher complexity of a class is critical in understanding the learnability of a problem. So how do we calculate the Rademacher complexity $\mathcal{R}^{\text{iid}}(\mathcal{F})$?

We start with a simple yet fundamental case: when $\mathcal{F}$ is finite. All subsequent discussions on infinite classes will eventually make use of the results for finite classes. The key lemma we need is the following so-called maximal inequality for *sub-Gaussian* random variables. Recall that a zero-mean random variable U is σ-sub-Gaussian if $\mathbb{E}[\exp(\lambda U)] \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda > 0$, that is, its moment generating function is bounded by that of a zero-mean Gaussian with variance σ^2. For example, any zero-mean random variable with range $[a, b]$ is $\frac{b-a}{2}$-sub-Gaussian (this is the so-called Hoeffding's lemma).

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

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

Proof. For any $\lambda > 0$, we have

$$\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\max_{f\in\mathcal{F}} U_f\right]\right) &\le \mathbb{E}\left[\exp\left(\lambda \max_{f\in\mathcal{F}} U_f\right)\right] && \text{(Jensen's inequality)}\\
&\le \mathbb{E}\left[\sum_{f\in\mathcal{F}} \exp\left(\lambda U_f\right)\right]\\
&\le \sum_{f\in\mathcal{F}} \exp\left(\sigma^2\lambda^2/2\right) && (U_f \text{ is } \sigma\text{-sub-Gaussian})\\
&= |\mathcal{F}| \exp\left(\sigma^2\lambda^2/2\right).
\end{aligned}$$

Rearranging gives $\mathbb{E}\left[\max_{f\in\mathcal{F}} U_f\right] \le \frac{\ln|\mathcal{F}|}{\lambda} + \frac{\sigma^2\lambda}{2}$, which is $\sigma\sqrt{2\ln|\mathcal{F}|}$ by setting $\lambda = \sqrt{2\ln|\mathcal{F}|}/\sigma$ (this choice of λ minimizes the upper bound). □

Next we apply this maximal inequality to bound the Rademacher complexity for a finite class.

Theorem 2 (Massart's Lemma). *Let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class, and $x_1, \ldots, x_n \in \mathcal{X}$ be an arbitrary set of inputs. We have*

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le \frac{1}{n}\sqrt{2\left(\max_{f\in\mathcal{F}} \sum_{t=1}^{n} f^2(x_t)\right)\ln|\mathcal{F}|}.$$

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

Proof. Note that $\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) = \frac{1}{n}\mathbb{E}\left[\max_{f\in\mathcal{F}} U_f\right]$ where $U_f = \sum_{t=1}^{n} \epsilon_t f(x_t)$. The following calculation shows that U_f is σ-sub-Gaussian with $\sigma = \max_{f\in\mathcal{F}} \sqrt{\sum_{t=1}^{n} f^2(x_t)}$: for any $\lambda > 0$

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda U_f)\right] &= \Pi_{t=1}^{n} \mathbb{E}\left[\exp(\lambda \epsilon_t f(x_t))\right] \le \Pi_{t=1}^{n} \exp\left(f^2(x_t)\lambda^2/2\right)\\
&= \exp\left(\left(\sum_{t=1}^{n} f^2(x_t)\right)\lambda^2/2\right) \le \exp\left(\sigma^2\lambda^2/2\right),
\end{aligned}$$

where the first step uses the fact that $\epsilon_1, \ldots, \epsilon_n$ are independent, and the second step uses the fact that $\epsilon_t f(x_t)$ is $|f(x_t)|$-sub-Gaussian. Applying Lemma 3 then finishes the proof. □

The theorem above implies that finite classes with bounded value in $[-C, C]$ are all learnable since $\mathcal{V}^{\text{iid}}(\mathcal{F}, n) \le 2G \sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\mathcal{F}) \le 2GC\sqrt{\frac{2\ln|\mathcal{F}|}{n}} \to 0$ as n goes to infinity. In fact, it also tells us the exact convergence rate ($1/\sqrt{n}$) when we learn via an ERM. One might notice that the same conclusion can be reached by applying the maximal inequality directly to the empirical process ($\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t)\right)\right]$), but very soon we will see why it is important to do so on the Rademacher complexity instead.

3 Infinite Class: Classification

We next move on to study the Rademacher complexity of infinite classes. As the first step, we consider binary classification problems with $\mathcal{Y} = \{-1, +1\}$. While Lemma 3 is seemingly not useful for infinite classes, it in fact also plays a key role here. The main observation is the following: the conditional Rademacher complexity can be equivalently written as

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t)\right] = \frac{1}{n}\mathbb{E}\left[\max_{v\in\mathcal{F}|_{x_{1:n}}} \sum_{t=1}^{n} \epsilon_t v_t\right]$$

where $\mathcal{F}|_{x_{1:n}} = \{(f(x_1), \cdots, f(x_n)) \mid f \in \mathcal{F}\} \subset \{-1, +1\}^n$ is the *projection* of $\mathcal{F}$ onto the input set $x_{1:n}$. While $\mathcal{F}$ is infinite, $\mathcal{F}|_{x_{1:n}}$ on the other hand is always finite! In particular, its size is never larger than 2^n. Based on this intuition, we define the *growth function* of a class $\mathcal{F}$ for n inputs as

$$\Pi_{\mathcal{F}}(n) = \max_{x_{1:n}} \left|\mathcal{F}|_{x_{1:n}}\right|,$$

which is the maximum number of labeling one can possibly obtain using functions from $\mathcal{F}$ for n samples. Based on the previous observation and Lemma 3, we immediately have

$$\mathcal{R}^{\text{iid}}(\mathcal{F}) \le \sqrt{\frac{2\ln \Pi_{\mathcal{F}}(n)}{n}}.$$

As mentioned, a trivial upper bound on the growth function $\Pi_{\mathcal{F}}(n)$ is 2^n, which, when plugged into the bound above, leads to a constant Rademacher complexity. Therefore, to hope for a vanishing Rademacher complexity, we require the class to have a much milder growth function. Below, we discuss two such examples.

Proposition 1. *Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \left\{ f_\theta(x) = \begin{cases} +1 & \text{if } x \le \theta \\ -1 & \text{else} \end{cases} \;\middle|\; \theta \in \mathbb{R} \right\}$ is the set of threshold functions. Then $\Pi_{\mathcal{F}}(n) = n+1$ and thus $\mathcal{V}^{\text{iid}}(\mathcal{F}, n) \le \sqrt{\frac{2\ln(n+1)}{n}}$.*

Proof. For any θ, $f_\theta(x)$ classifies all the points to the left of θ as $+1$ and all the points to the right as -1. Clearly, for any n distinct points on the real line, all the $n+1$ possible labelings are (from left to right): $\{-1,-1,\ldots,-1\}, \{+1,-1,\ldots,-1\}, \{+1,+1,\ldots,-1\}, \cdots, \{+1,+1,\ldots,+1\}$. □

Proposition 2. *Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \left\{ f_{\theta_1,\theta_2}(x) = \begin{cases} +1 & \text{if } \theta_1 \le x \le \theta_2 \\ -1 & \text{else} \end{cases} \;\middle|\; \theta_1 \le \theta_2 \right\}$ is the set of interval functions. Then $\Pi_{\mathcal{F}}(n) = \binom{n+1}{2} + 1 = \mathcal{O}(n^2)$ and thus $\mathcal{V}^{\text{iid}}(\mathcal{F}, n) \le \mathcal{O}\left(\sqrt{\frac{\ln n}{n}}\right)$.*

Proof. Any n distinct points divide the real line into $n+1$ regions. Putting the interval endpoints θ_1 and θ_2 into any two of these regions gives $\binom{n+1}{2}$ labelings. Putting the interval endpoints into the same region (any one of them) will give one extra labeling with all -1 labels. □

We remark that while the step of replacing $\sup_{f\in\mathcal{F}}$ with $\max_{v\in\mathcal{F}|_{x_{1:n}}}$ in the definition of Rademacher complexity is very intuitive and straightforward, one in fact cannot do the same thing directly for $\mathcal{V}^{\text{iid}}(\mathcal{F}, n)$ or its upper bound $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(L(f) - \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)\right)\right]$, because the term $L(f)$ depends on f not just through the values $f(x_1), \ldots, f(x_n)$. This highlights the importance of relaxing these quantities to the Rademacher complexity via symmetrization, and why we did not directly apply the maximal inequality to the empirical process for a finite class.

3.1 VC dimension and Sauer's lemma

While the growth function is a nice way to characterize the complexity of a class, it is not always easy to compute. To see this, consider an example where $\mathcal{X} = \mathbb{R}^d$ for some dimension d and

$$\mathcal{F} = \left\{ f_{\theta,b}(x) = \text{sign}\left(\langle x, \theta\rangle + b\right) \mid \theta \in \mathbb{R}^d, b \in \mathbb{R} \right\} \tag{3}$$

is the set of linear classifiers (sign(y) is $+1$ if $y \ge 0$ and -1 otherwise). What is $\Pi_{\mathcal{F}}(n)$ in this case? For simplicity let's start from $d = 2$. It is pretty clear that for any $n \le 3$, one can find n points so that $\mathcal{F}$ realizes all the possible 2^n labelings, and thus we know $\Pi_{\mathcal{F}}(n) = 2^n$ for $n \le 3$. What about $n = 4$? Well, first we know $\Pi_{\mathcal{F}}(4) < 2^n$ because for any four points in a 2D plane, it is impossible for linear classifiers to realize all the 16 possible labelings (try to convince yourself). But what exactly is the value of $\Pi_{\mathcal{F}}(4)$? After spending some time you probably can figure this out as well. But what about $\Pi_{\mathcal{F}}(5)$, $\Pi_{\mathcal{F}}(6)$, and more generally $\Pi_{\mathcal{F}}(n)$ for an arbitrary n? Do we need to figure out all these values, which appears to be a tedious process?

Somewhat surprisingly, it turns out that the two facts we mentioned above: $\Pi_{\mathcal{F}}(3) = 2^3$ and $\Pi_{\mathcal{F}}(4) < 2^4$, are already enough to derive a pretty tight upper bound on $\Pi_{\mathcal{F}}(n)$ for an arbitrary n! To show this result, we first make a few definitions. We say that $\mathcal{F}$ *shatters* a set of inputs $x_{1:n}$ if $\mathcal{F}|_{x_{1:n}} = \{-1,+1\}^n$, that is, $\mathcal{F}$ realizes all the 2^n possible labelings of this set. The *Vapnik–Chervonenkis* (VC) dimension of $\mathcal{F}$ is defined as the size of the largest input set that can be shattered by $\mathcal{F}$, that is,

$$\text{VCdim}(\mathcal{F}) = \max\left\{ n : \Pi_{\mathcal{F}}(n) = 2^n \right\}.$$

If the set is empty, then $\text{VCdim}(\mathcal{F})$ is defined as 0; and if the set is not finite (so $\Pi_{\mathcal{F}}(n) = 2^n$ for all n), $\text{VCdim}(\mathcal{F})$ is defined as ∞. As an example, the VC-dimension of the 2D linear classifiers discussed above is 3. The following seminal result connects the growth function and the VC-dimension of a class (proof is deferred to the next subsection).

Lemma 4 (Sauer's lemma)**.** *For a class $\mathcal{F}$ with finite VC-dimension d, one has for any $n > d$*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d .$$

By the definition of VC-dimension, we clearly have $\Pi_{\mathcal{F}}(n) = 2^n$ for any $n \leq d$. What Sauer's lemma shows is that once n becomes larger than d, there is a phase transition and the exponential growth (2^n) for the growth function suddenly becomes a polynomial growth (roughly n^d)! Combing with the previous discussion, we thus have

$$\mathcal{R}^{\text{iid}}(\mathcal{F}) \leq \sqrt{\frac{2\ln \Pi_{\mathcal{F}}(n)}{n}} \leq \sqrt{\frac{2d\ln\left(\frac{en}{d}\right)}{n}},$$

implying that a class with finite VC-dimension is always learnable.

We emphasize that to prove $\text{VCdim}(\mathcal{F}) = d$, one needs to show exactly two things: 1) $\Pi_{\mathcal{F}}(d) = 2^d$, that is, *provide a concrete set* of inputs of size d and prove that $\mathcal{F}$ realizes all possible labelings on this set; and 2) $\Pi_{\mathcal{F}}(d+1) < 2^{d+1}$, that is, prove that for *any* input set of size $d+1$, there exists a labeling that is not achievable by $\mathcal{F}$. This is often easier to do compared to finding the growth function for any n, as we already see for the linear classifier example. Below are a few more examples.

Proposition 3. *A class has VC-dimension* 0 *if and only if it contains only one function.*

Proposition 4. *The threshold function class defined in Proposition 1 has VC-dimension* 1.

Proposition 5. *The interval function class defined in Proposition 2 has VC-dimension* 2.

You should be able to prove these statements without too much difficulty. For threshold and interval function class, we figured out the exact growth function earlier, and one can see that the upper bound given by Sauer's lemma is very tight.

Proposition 6. *The linear classifier class defined in Equation (3) has VC-dimension* $d+1$.

We have proved this statement for $d = 2$ in earlier discussion. Proving the general case will be in HW 1. By now, you might notice that the VC-dimension often matches the number of parameters of the class. Indeed, this often serves as a quick (and most of the time, accurate) guess on the VC-dimension. This is, however, not always correct, as shown in the following example where a class with a single parameter has infinite VC-dimension. The intuition is that by picking a large enough θ, the function $\sin(\theta x)$ can wiggle arbitrarily often within a small interval (see HW1).

Proposition 7. *Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_\theta(x) = \text{sign}(\sin(\theta x)) \mid \theta \in \mathbb{R}\}$. Then* $\text{VCdim}(\mathcal{F}) = \infty$.

3.2 Proof of Sauer's lemma

Proof. We prove the statement $\Pi_{\mathcal{F}}(n) \leq g(d,n) = \sum_{i=0}^{d}\binom{n}{i}$ for $n > d$ by induction on the value of $d+n$. The base case $d+n=1$ is trivial: the only possible configuration is $d=0$ and $n=1$, and in this case $\Pi_{\mathcal{F}}(n) = 1 = g(0,1)$. Next we assume that the statement holds for any $n' > d'$ such that $n'+d' < n+d$, and prove $\Pi_{\mathcal{F}}(n) \leq g(d,n)$. The case when $d=0$ is again trivial, so we assume $n > d > 0$. For any set of distinct inputs $x_{1:n}$, let $F_1 = \mathcal{F}|_{x_{2:n}}$ be the projection of $\mathcal{F}$ onto the $n-1$ inputs $x_{2:n}$ and $F_2 \subset F_1$ be such that

$$F_2 = \left\{v \in F_1 \mid (-1,v),(+1,v) \in \mathcal{F}|_{x_{1:n}}\right\},$$

that is, for any labeling in F_2 for $x_{2:n}$, adding x_1 with either label leads to a labeling of $x_{1:n}$ that can be realized by $\mathcal{F}$. It is clear that

$$\left|\mathcal{F}|_{x_{1:n}}\right| = |F_1| + |F_2|.$$

Now, we see F_1 and F_2 as two function classes defined only on $x_{2:n}$ (so a function is just a vector in $\{-1,+1\}^{n-1}$). Then clearly we have

$$|F_1| = \left|F_1|_{x_{2:n}}\right| \leq \Pi_{F_1}(n-1) \leq g(d,n-1),$$

where the last step uses the inductive hypothesis and the fact the F_1 cannot have VC-dimension larger than that of $\mathcal{F}$. On the other hand, we also have

$$|F_2| = \left|F_2|_{x_{2:n}}\right| \le \Pi_{F_2}(n-1) \le g(d-1, n-1),$$

where the last step uses the inductive hypothesis and the fact the F_2 has VC-dimension at most $d-1$, since otherwise, there exists a subset of $x_{2:n}$ of size d that can be shattered by F_2, and adding x_1 to this subset leads to a set of size $d+1$ that can be shattered by $\mathcal{F}$ due to the construction of F_2, which is a contradiction to the condition VCdim$(\mathcal{F}) = d$. Together this implies

$$\begin{aligned}
\left|\mathcal{F}|_{x_{1:n}}\right| &\le g(d, n-1) + g(d-1, n-1) \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=1}^{d} \binom{n-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{n}{i} = g(d, n).
\end{aligned}$$

Since this holds for any $x_{1:n}$, we thus have $\Pi_{\mathcal{F}}(n) \le g(d, n)$, finishing the inductive proof. The second statement of the inequality holds because

$$\begin{aligned}
g(d, n) = \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \left(\frac{d}{n}\right)^d \binom{n}{i} &\le \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \left(\frac{d}{n}\right)^i \binom{n}{i} && (d < n) \\
\le \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \left(\frac{d}{n}\right)^i \binom{n}{i} &= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \le \left(\frac{en}{d}\right)^d. && (1 + x \le e^x,\ \forall x \in \mathbb{R})
\end{aligned}$$

□

4 Summary and Closing the Loop

This lecture can be summarized by the following sequence of upper bounding:

$$\begin{aligned}
\mathcal{V}^{\text{iid}}(\mathcal{F}, n) &\le \sup_{\mathcal{P}} \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(L(f) - \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right) \right] \right) && \text{(using ERM)} \\
&\le 2 \sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\ell(\mathcal{F})) && \text{(symmetrization)} \\
&\le 2G \sup_{\mathcal{P}} \mathcal{R}^{\text{iid}}(\mathcal{F}) && \text{(erasing the loss)} \\
&\le \begin{cases} 2GC\sqrt{\frac{2\ln|\mathcal{F}|}{n}} & \text{(finite class)} \\ \sqrt{\frac{2\ln \Pi_{\mathcal{F}}(n)}{n}} \le \sqrt{\frac{2d\ln\left(\frac{en}{d}\right)}{n}}, & \text{(binary classification)} \end{cases}
\end{aligned}$$

where again G is $1/2$ for a binary classification problem or the Lipschitz constant for the regression loss, C is a bound on the magnitude of the function value, and $d = \text{VCdim}(\mathcal{F})$. In the end, we found that for binary classification, having a finite VC-dimension is a sufficient condition for learnability, but is it also necessary, or in other words, is this sequence of upper bounding tight enough?

The answer is yes: a finite VC-dimension is also necessary for learnability, so we basically have a closed loop. Indeed, if a class $\mathcal{F}$ has an infinite VC-dimension, then for any n, we can find a subset $\mathcal{X}' \subset \mathcal{X}$ with $2n$ elements shattered by $\mathcal{F}$, that is, $\mathcal{F}$ behaves the same as $\mathcal{Y}^{\mathcal{X}'}$ on this set. Therefore, by the exact same argument of the no free lunch theorem discussed in Lecture 1, for any algorithm one can find a distribution $\mathcal{P}$ supported on $\mathcal{X}' \times \mathcal{Y}$ such that it suffers excess risk at least $1/4$, implying that $\mathcal{F}$ is not learnable.

Note that this also implies that if a class is learnable (for a binary classification problem), then it must be learnable via the simple ERM algorithm. As a final remark, we mention (without going into details) that this is not always the case for general statistical learning.