Topics in Deep Learning Theory (Spring 2025)

Lecture 1: Generalization via Uniform Concentration

Instructor: Lei Wu Date: May 16, 2025

1 The framework of statistical learning

In supervised learning, the objective is to learn a target $h^* : \mathcal{X} \mapsto \mathcal{Y}$ using a finite set of samples $S_n = \{(x_i, y_i)\}$ where each $y_i = h^*(x_i) + \xi_i$ represents a potentially noisy measurement of $h^*(x_i)$. A fundamental question in learning theory is:

Question 1: *How many samples are required to learn f^* effectively?*

Statistical learning theory addresses this question by framing it within a probabilistic context. Assume there exist an underlying probability distribution $\mu \in \mathcal{P}(\mathcal{X})$ and $x_1, x_2, \ldots, x_n$ are i.i.d. samples drawn from μ. Given a learned model $\hat{h} : \mathcal{X} \mapsto \mathcal{Y}$ (which are often learned using S_n), its performance is evaluated using the expected loss under ρ:

$$\mathcal{E}(\hat{h}; h^*) := \mathbb{E}_{x\sim\mu}[\ell(\hat{h}(x), h^*(x))]. \tag{1}$$

This quantity, known as the *generalization error*, measures how well $\hat{h}$ approximates h^* on average over μ.

Remark 1.1. *It should be remarked that in practice, $x_1, \ldots, x_n$ are not always be i.i.d. and the performance might be assessed using metrics different from (1). Nevertheless, the above idealized setup provides a reasonable and analytically tractable scenarios for answering Question 1 in a quantitative way.*

To make the problem more manageable, statistical learning often shifts focus to a broader, worst-case question:

Given a function class (or hypothesis set) $\mathcal{H}$, how many samples are required to learn functions within $\mathcal{H}$? How does this depend on the complexity of $\mathcal{H}$?

This formulation aligns with the technique of establishing a uniform bound over the entire class:

$$\mathcal{E}(\hat{h}; h^*) \le \sup_{h^*\in\mathcal{H}} \mathcal{E}(\hat{h}; h^*). \tag{2}$$

This worst-case approach ensures robust performance across all possible target functions in $\mathcal{H}$. However, the sample complexity for learning a specific $h^* \in \mathcal{H}$ can sometimes be significantly lower than that required for uniform learning over the entire class $\mathcal{H}$. A tighter estimate may be achieved by identifying a smaller subclass $\mathcal{H}' \subset \mathcal{H}$ such that $h^* \in \mathcal{H}'$, thereby reducing the complexity.

In this lecture, we will demonstrate why the worst-case approach makes the analysis easier and how to bound the uniform excess risk in (2) by leveraging two fundamental tools from probability theory:

- **Concentration inequalities**: These inequalties quantify the rate at which the empirical mean converges to the population mean, offering probabilistic bounds on the deviations between the two.
- **Maximal inequalities**: These inequalties enable the uniform bound step in (2) by controlling the supremum over the hypothesis class $\mathcal{H}$. Since $\mathcal{H}$ may be large or even infinite, accurately measuring its complexity is critical for deriving tight bounds. This section focuses on developing such measures, including covering numbers and Rademacher complexity.

2 Setup

Let $z = (x, y), \ell_h(z) = \ell(h(x), y)$, and

$$\begin{aligned} \hat{\mathcal{R}}(h) &= \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i), \\ \mathcal{R}(h) &= \mathbb{E}_z[\ell_h(z)] \end{aligned} \tag{3}$$

be the empirical risk and population risk, respectively. Let $\mathcal{H}$ be a hypothesis class. Consider the estimator:

$$\hat{h}_n = \operatorname*{argmin}_{h\in\mathcal{H}} \hat{\mathcal{R}}(h).$$

This type of estimator ensures the smallness of $\hat{\mathcal{R}}(\hat{h}_n)$, but how small is $\mathcal{R}(\hat{h}_n)$?

For any $h \in \mathcal{H}$, consider the decomposition:

$$\mathcal{R}(h) = \underbrace{\hat{\mathcal{R}}(h)}_{\text{training error}} + \underbrace{\mathcal{R}(h) - \hat{\mathcal{R}}(h)}_{\text{gen-gap}},$$

where the generalization gap satisfies

$$\text{gen-gap}(h) := \mathcal{R}(h) - \hat{\mathcal{R}}(h) = \mathbb{E}_z[\ell_h(z)] - \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i). \tag{4}$$

One may expect that $\text{gen-gap}(h) = O(1/\sqrt{n})$. By concentration inequality, this is true for h that is independent of training data $(z_1, \ldots, z_n)$. However, our task is to bound the gen-gap of $\hat{h}_n$:

$$\text{gen-gap}(\hat{h}_n) = \mathbb{E}_z[\ell_{\hat{h}_n}(z)] - \frac{1}{n}\sum \ell_{\hat{h}_n}(z_i).$$

Note that $\hat{h}_n$ depends on $(z_1, \ldots, z_n)$ and hence $\{\ell_{\hat{h}_n}(z_i)\}$ are not i.i.d. . Consequently, gen-gap may not be in the order of $O(1/\sqrt{n})$. In fact, $\text{gen-gap}(\hat{h}_n)$ can be arbitrarily large if $\hat{h}_n$ is a very complex solution.

3 Uniform bounds

To address the issue of dependence, we consider the uniform bound:

$$\left|\mathcal{R}(\hat{h}_n) - \hat{\mathcal{R}}(\hat{h}_n)\right| \le \sup_{h\in\mathcal{H}} \left|\mathcal{R}(h) - \hat{\mathcal{R}}(h)\right|. \tag{5}$$

Clearly, when the hypothesis space $\mathcal{H}$ is sufficiently "small"—for instance, in the extreme case where $\mathcal{H} = \{h\}$—it is expected that

$$\sup_{h\in\mathcal{H}} \left|\mathcal{R}(h) - \hat{\mathcal{R}}(h)\right| \sim \frac{1}{\sqrt{n}}.$$

This raises several natural questions:

- What conditions on $\mathcal{H}$ ensure that the uniform bound remains small?
- What is the corresponding rate? Do we still obtain $O(1/\sqrt{n})$? Can the rate be faster or slower?

Before delving into the technical details, it is helpful to develop some intuition by considering specific learning problems:

- **Constant Function.** Consider the setting where $y_i = 1 + \xi_i$ with $\xi_i \sim \mathcal{N}(0, \sigma^2)$ and the true function $h^*(x) \equiv 1$. Here, the learning problem reduces to estimating the mean of a Gaussian distribution. Clearly, when $\sigma \gtrsim 1$, the rate of learning h^* (i.e., estimating the mean) cannot be faster than $O(n^{-1/2})$. This suggests that in the presence of non-negligible noise, the learning rate cannot exceed the standard Monte Carlo rate $O(n^{-1/2})$, which is a well-known result in statistical learning theory.
- **Noiseless Regime.** When $\sigma = 0$, learning can be arbitrarily fast. Consider the problem of learning a quadratic function $h^*(x) = x^2$. If the labels are noiseless, learning can proceed at a rate faster than the standard Monte Carlo rate (e.g., you can do piecewise linear interpolation.). Notably, in modern machine learning, many datasets are generated via high-fidelity numerical simulations, for which this noiseless scenario is highly relevant. However, the sample complexity of learning in this setting is still not well understood.

  We focus on the traditional statistical learning framework, where noise is assumed to be at a constant level. Consequently, we expect that the learning rate cannot surpass the Monte Carlo rate.

To build a more general intuition, let us first examine a simple case: learning with a finite hypothesis class.

Lemma 3.1 (Finite class)**.** *Let $\mathcal{H}$ be a collection of finite hypotheses and denote by $|\mathcal{H}|$ the number of hypotheses. Assume $\sup_{y,y'} |\ell(y, y')| \le 1$. For any $\delta \in (0, 1)$, with probability $1 - \delta$ over the random sampling of the training set S, we have*

$$\sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \le \sqrt{\frac{2\ln(2|\mathcal{H}|/\delta)}{n}}.$$

Proof. WLOG, suppose $\mathcal{H} = \{h_1, \ldots, h_m\}$. Let $z = (x, y)$ and $Q_h(z) = \ell(h(x), y)$. Taking the union bound gives us

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} |\frac{1}{n}\sum_{i=1}^{n} Q(h, z_i) - \mathbb{E}_z[Q(h, z)]| \ge t\right\} \le \sum_{j=1}^{m} \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Q(h_j, z_i) - \mathbb{E}_z[Z(h_j, z)]\right| \ge t\right\} \tag{6}$$

$$\leq m2e^{\frac{-2nt^2}{2^2}} = 2me^{\frac{-nt^2}{2}}, \tag{7}$$

where the last step follows from the Hoeffding's inequality. Let the failure probability $2me^{\frac{-nt^2}{2}} = \delta$, which leads to $t = \sqrt{\frac{2\ln(2m/\delta)}{n}}$.

□

Note that this lemma follows trivially from the maximal inequality, and its proof essentially follows the same steps as the proof of the maximal inequality. We include it here for completeness due to its simplicity.

We see that the upper bound only depends on the cardinality of hypothesis class $|\mathcal{H}|$ logarithmically. This implies that even when the hypothesis class has exponentially many functions, the generalization gap can still be well controlled.

Implication for quantized models. Consider a general model that has m parameters and all parameters are represented using k-bit floating-point number. Then, this model can represent 2^{km} functions. Consequently, the corresponding generalization gap must be bounded by $\sqrt{\frac{km+\log(1/\delta)}{n}}$. This means, in such a general case, the number of parameters is a good parameter to bound generalization. Unfortunately, the generalization is guaranteed for the underparameterized regime.

Definition 3.2 (Empirical process)**.** Let $\mathcal{F}$ be a class of real-valued functions $f : \Omega \mapsto \mathbb{R}$ where (Ω, Σ, μ) is a probability space. Let $X \sim \mu$ and $X_1, \ldots, X_n$ be independent copies of X. Then, the random process $(\mathbb{X}_f)_{f\in\mathcal{F}}$ defined by

$$\mathbb{X}_f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)$$

is called an *empirical process* indexed by $\mathcal{F}$.

With an abuse of notation, in our case, $f(Z) = \ell(h(X), Y)$. Our task is to bound the supremum of empirical process:

$$\sup_{f\in\mathcal{F}} \mathbb{X}_f.$$

We emphasize that bounding the supremum of a stochastic process is a classical problem in probability theory. To better understand this problem, let's start by comparing it with the vanilla maximal inequality. Recall that for n sub-Gaussian random variables $X_1, \ldots, X_n$, each with a variance proxy σ^2, the maximal inequality provides a clear bound:

$$\mathbb{E}\left[\sup_{i\in[n]} X_i\right] \leq \sigma\sqrt{2\log n}.$$

Similarly, if each $\mathbb{X}_f$ in a function class $\mathcal{F}$ is sub-Gaussian with a variance proxy bounded by σ^2, we might anticipate an analogous result:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{X}_f\right] \leq \sigma\sqrt{2\log|\mathcal{F}|}.$$

Using Hoeffding's or Chernoff's inequalities, we can demonstrate that $\mathbb{X}_f$ is sub-Gaussian with $\sigma^2 \leq C/n$. Substituting this into the previous bound yields:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{X}_f\right] \leq \sqrt{\frac{2C\log|\mathcal{F}|}{n}},$$

where C is a constant. This derivation holds firmly when $\mathcal{F}$ is finite, as $\log|\mathcal{F}|$ remains well-defined. However, when $\mathcal{F}$ becomes infinite, $\log|\mathcal{F}|$ grows unbounded, making the inequality vacuous. To address infinite classes, we need a more refined interpretation of $\log|\mathcal{F}|$. Fortunately, the tools introduced in this lecture note can be adapted to tackle such general scenarios, including infinite function classes. For a more comprehensive treatment, interested readers may explore [Vershynin, 2018].

4 Covering number or metric entropy

For the finite hypothesis classes, we have shown that $\log|\mathcal{F}|$, i.e., the logarithm of cardinality, can be used as a good complexity measure. However, there are two major problems with using the $|\mathcal{F}|$ as the complexity measure:

- It does not apply to the case with $|\mathcal{F}| = \infty$, which is common in practice.
- It does not exploit any structure of $\mathcal{F}$ and arguably, we should be able to reduce the number of samples required by utilizing these structures.

One possible property that we can use is: $\mathcal{F}$ is **compressible**. One possible approach is *discretization*. This means that we choose a finite subset $\mathcal{F}_\varepsilon \subset \mathcal{F}$ to "represent" $\mathcal{F}$.

Definition 4.1 (Covering number)**.** Consider a metric space (T, ρ).

- We say $T_\varepsilon \subset T$ is an ε-cover (also called ε-net) of T, if for any $t \in T$, there exists a $t' \in T_\varepsilon$ such that $\rho(t, t') \leq \varepsilon$.
- The covering number $\mathcal{N}(T, \rho, \varepsilon)$ is defined as the smallest cardinality of an ε-cover of T with respect to ρ.

Definition 4.2 (Metric entropy)**.** The *metric entropy* of T is defined by $\log \mathcal{N}(T, \rho, \varepsilon)$.

Remark 4.3. *What does $\log_2 \mathcal{N}(T, \rho, \varepsilon)$ represent? It denotes the number of bits required to compress the set T up to a resolution ϵ under the metric ρ. In this context, the metric entropy, defined as $\log_2 \mathcal{N}(\mathcal{F}, \rho, \epsilon)$, quantifies the compressibility of a function class and, consequently, provides insight into the complexity of a model.*

Theorem 4.4. *Let $\mathcal{F}$ be a function class with $\sup_{f\in\mathcal{F}, x\in\mathcal{X}} |f(x)| \leq B$. Let $\|f - g\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - g(x)|$. Then, for any $\delta \in (0, 1)$, w.p. at least $1 - \delta$ over the sampling of $X_1, X_2, \ldots, X_n$, we have*

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \leq 2\varepsilon + B\sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) + \log(2/\delta)}{n}}.$$

Proof. Let $\mathcal{F}_\varepsilon$ be an ε-cover of $\mathcal{F}$. For any $f \in \mathcal{F}$, let $f' \in \mathcal{F}_\varepsilon$ such that $\|f - f'\|_\infty \le \varepsilon$. Then, we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| \le& \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f'(X_i)\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^n f'(X_i) - \mathbb{E}[f'(X)]\right| + \left|\mathbb{E}\, f'(X) - \mathbb{E}[f(X)]\right|.\end{aligned}$$

Taking the surprimum with respect to $f \in \mathcal{F}$ gives

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| &\le 2\varepsilon + \sup_{f'\in\mathcal{F}_\varepsilon}\left|\frac{1}{n}\sum_{i=1}^n f'(X_i) - \mathbb{E}[f'(X)]\right| \\ &\le 2\varepsilon + 2B\sqrt{\frac{\log(2|\mathcal{F}_\varepsilon|/\delta)}{n}},\end{aligned}$$

where the last step uses Lemma 3.1. By definition, for any $q > 0$, there exist an ϵ-cover $\mathcal{F}_{\epsilon,q}$ such that $|\mathcal{F}_{\varepsilon,q}| \le \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) + q$. Thus, choosing the $\mathcal{F}_\epsilon$ to be $\mathcal{F}_{\epsilon,q}$ and taking $q \to 0$, we complete the proof. □

We note that Theorem 4.4 is fairly intuitive but not sufficiently tight for two reasons:

- The metric entropy should ideally be defined with respect to an average-like metric rather than the supremum norm. This limitation can be addressed using the symmetrization technique introduced later.
- Discretization at a single resolution is inadequate; this issue will be resolved by employing the multi-resolution analysis via chaining.

Compressibility implies learnability. This theorem suggests that the metric entropy of a function class (and its associated model) governs its learnability. Additionally, as noted in Remark 4.3, metric entropy also quantifies the compressibility of the function class. Together, these insights lead to an intriguing conclusion:

If a model is compressible, it is also learnable.

Thus, metric entropy serves as a bridge linking compressibility to learnability.

Example: Lipschtiz models. Let $f : \mathcal{X} \times \mathbb{R}^m \mapsto \mathbb{R}$ be our model, where m denotes the number of parameters. Assume that f is L-Lipschitz in the sense that $\sup_x |f(x;\theta_1) - f(x;\theta_2)| \le L\rho(\theta_1, \theta_2)$.

Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Omega\}$ be the function class. Let Ω_ε be an ε-cover of Ω with respect to the ρ metric. Then,

$$\|f(\cdot;\theta_1) - f(\cdot;\theta_2)\|_\infty \le L\rho(\theta_1, \theta_2).$$

implies that $\mathcal{F}_\varepsilon = \{f(\cdot;\theta) : \theta \in \Omega_{\varepsilon/L}\}$ is an ε-cover of $\mathcal{F}$. Hence, we have

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(\Omega, \rho, \varepsilon/L). \tag{8}$$

Linear class. Consider the linear class:

$$\mathcal{H} = \left\{ x \mapsto w^\top x : \|w\|_2 \le 1, \|x\|_2 \le 1 \right\}.$$

Then,

$$\sup_{\|x\|_2 \le 1} |w^\top x - v^\top x| \le \|w - v\|_2 \sup_{\|x\|_2 \le 1} \|x\|_2 \le \|w - v\|_2.$$

Let $B_d(r) = \{x \in \mathbb{R}^d : \|x\| \le r\}$ be the ball of radius r. Then, (8) gives

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(B_d(1), \|\cdot\|_2, \varepsilon).$$

The above examples demonstrate that one can reduce the estimation of covering number of a function class to the covering number of a subset in Euclidean space. The latter is often easier to estimate and we provide below one of the most important examples.

4.1 Volume argument

To help the estimation of covering number, we introduce the packing number.

Definition 4.5 (Packing number)**.** Consider a metric space (T, ρ). $T_\varepsilon \subset T$ is said to be ε-separated if $\rho(x, y) > \varepsilon$ for any $x, y \in T_\varepsilon$ and $x \neq y$. The packing number is defined as

$$\mathcal{P}(\mathcal{F}, \rho, \varepsilon) = \sup_{T_\varepsilon \subset T \text{ is } \varepsilon\text{-separated}} |T_\varepsilon|.$$

Lemma 4.6. $\mathcal{N}(T, \rho, \varepsilon) \le \mathcal{P}(T, \rho, \varepsilon)$.

Proof. Let T_ε be the maximal ε-separated subset. Then, we claim that T_ε is also an ε-cover of T, i.e., $T \subset \cup_{x \in T_\varepsilon} B(x; \varepsilon)$. If not, there exists a $y \in T$ such that $d(y, x) > \varepsilon$ for any $x \in T_\varepsilon$. Hence, $T_\varepsilon \cup \{y\}$ is also ε-separated, which is contradictory with the assumption. □

Lemma 4.7. $(1/\varepsilon)^d \le \mathcal{N}(B_d(1), \|\cdot\|_2, \varepsilon) \le (1 + 2/\varepsilon)^d$.

The proof follows from a volume argument.

Proof. **Lower bound.** Let N_ε be an ε-cover of $B_d(1)$. Then, $B_d(1) \subset \cup_{x \in N_\varepsilon} B_d(x; \varepsilon)$. Therefore,

$$\mathrm{Vol}(B_d(1)) \le \sum_{x \in N_\varepsilon} \mathrm{Vol}(B_d(x; \varepsilon)) = |N_\varepsilon| \mathrm{Vol}(B_d(x; \varepsilon)).$$

Hence,

$$\mathcal{N}(B_d(1), \|\cdot\|_2, \varepsilon) = |N_\varepsilon| \ge \frac{\mathrm{Vol}(B_d(1))}{\mathrm{Vol}(B_d(x; \varepsilon))} = \left(\frac{1}{\varepsilon}\right)^d.$$

Upper bound. Let $P_\varepsilon \subset B_d(1)$ be ε-separated. Then, by definition of packing number, we have

$$\cup_{x \in P_\varepsilon} B_d(x; \varepsilon/2) \subset B_d(1 + \varepsilon/2) \Rightarrow \sum_{x \in P_\varepsilon} \mathrm{Vol}(B_d(x; \varepsilon/2)) \le \mathrm{Vol}(B_d(1 + \varepsilon/2)).$$

Let $C_d r^d$ be the volume of a ℓ_2 ball of radius r. Then,

$$|P_\varepsilon| C_d (\varepsilon/2)^d \le C_d (1 + \varepsilon/2)^d \Rightarrow |P_\varepsilon| \le (1 + 2/\varepsilon)^d.$$

Then, the upper bound follows from Lemma 4.6. □

Remark 4.8. *The volume argument described above can also be utilized to estimate the covering number of other classes and under different metrics.*

5 Rademacher complexity

A random variable (R.V.) X is said to be symmetric if $-X \overset{d}{=} X$, where $\overset{d}{=}$ denotes equality in distribution. A key property of a symmetric R.V. is that its expectation is zero, i.e., $\mathbb{E}[X] = 0$. This property is especially valuable in problems where the tail behavior of the random variable is the primary focus. Symmetrization, a technique that leverages this symmetry, can often simplify the analysis of such problems by centering the variable while maintaining its tail characteristics, as we will explore further below.

Symmetrization of a Random Variable Let ξ be a symmetric R.V., meaning $-\xi \overset{d}{=} \xi$. Define $Z := \xi X$. Then, Z is also a symmetric R.V. This follows because:

$$-Z = (-\xi)X \overset{d}{=} \xi X = Z, \tag{9}$$

using the symmetry of ξ. A significant consequence of this symmetrization is that $\mathbb{E}[Z] = 0$. Moreover, the tail behavior of X is preserved in Z, meaning that Z and X exhibit similar tail characteristics. This preservation makes Z a useful tool for studying the properties of X.

Symmetrization of Empirical Processes Analogous to the symmetrization of a random variable as presented in equation (9), we introduce the following symmetrization for empirical processes:

$$\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)_{f\in\mathcal{F}} \longrightarrow \left(\frac{1}{n}\sum_{i=1}^{n} \xi_i f(X_i)\right)_{f\in\mathcal{F}},$$

where $\xi_1, \ldots, \xi_n$ are independent and identically distributed (i.i.d.) symmetric random variables. In particular, we focus on the case where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables, defined such that $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$. The lemma below demonstrates that this symmetrization does not significantly alter the supremum of the process.

Lemma 5.1 (Symmetrization of Empirical Processes)**.**

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right)\right] \le 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i f(X_i)\right)\right],$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables with $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$.

Proof. Let X_i' be an independent copy of X_i. To simplify the notation, we use $\mathbb{E}_{X_i}$ and $\mathbb{E}_{X_i'}$ to denote the expectation with respect to $\{X_i\}_{i=1}^n$ and $\{X_i'\}_{i=1}^n$, respectively. Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}}[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)] = \mathbb{E}_{X_i}\sup_{f\in\mathcal{F}}\mathbb{E}_{X_i'}[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(X_i'))] \tag{10}$$

$$\le \mathbb{E}_{X_i, X_i'}\sup_{f\in\mathcal{F}}[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(X_i'))]. \tag{11}$$

Due to that $f(X_i) - f(X_i')$ is symmetric [1], for any $\{\xi_i\} \in \{\pm 1\}^n$, we have

$$\begin{aligned}
\mathbb{E}_{X_i, X_i'} \sup_{f \in \mathcal{F}} [\frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X_i')] &= \mathbb{E}_{X_i, X_i'} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i [f(X_i) - f(X_i')] \\
&= \mathbb{E}_{X_i, X_i', \xi} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i [f(X_i) - f(X_i')] \\
&\leq \mathbb{E}_{X_i, X_i', \xi} [\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(X_i) + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\xi_i f(X_i')] \\
&= 2\, \mathbb{E}_{X_i, \xi} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(X_i).
\end{aligned}$$

□

Definition 5.2 (Rademacher complexity)**.** The empirical Rademacher complexity of a function class $\mathcal{F}$ on a set of training samples $\{x_i\}_{i=1}^n$ is defined as

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_\xi [\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i)].$$

The population Rademacher complexity is given by

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}[\widehat{\mathrm{Rad}}_n(\mathcal{F})],$$

where the expectation is taken over the distribution of $\{x_i\}_{i=1}^n$.

Thus, the symmetrization lemma (Lemma 5.1) can be restated as follows

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X) \right] \leq 2\, \mathrm{Rad}_n(\mathcal{F}). \tag{12}$$

This implies that the Rademacher complexity reflects the degree of concentration.

Theorem 5.3. *Assume that $0 \leq f \leq B$ for all $f \in \mathcal{F}$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the choice of the training set $S = \{X_1, \ldots, X_n\}$, we have*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq 2\, \mathrm{Rad}_n(\mathcal{F}) + B \sqrt{\frac{2 \log(2/\delta)}{n}}, \tag{13}$$

and the sample-dependent version:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq 2 \widehat{\mathrm{Rad}}_n(\mathcal{F}) + 4B \sqrt{\frac{2 \log(4/\delta)}{n}}. \tag{14}$$

[1] A random variable Z is said to be symmetric if Z and $-Z$ have the same distribution.

Proof. Let

$$G\left(x_1, \ldots, x_n\right) = \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} f\left(x_i\right) - \mathbb{E} f(X) \right].$$

Note that for any $i \in [n]$, it holds that

$$\begin{aligned}
&G(X_1, \ldots, X_n) - G(\tilde{X}_1, \ldots, \tilde{X}_n) \\
&\qquad = \sup_{f \in \mathcal{F}} (\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)) - \sup_{f \in \mathcal{F}} (\frac{1}{n} \sum_{i=1}^{n} f(\tilde{X}_i) - \mathbb{E}\, f(X)) \\
&\qquad \leq \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X) - \left(\frac{1}{n} \sum_{i=1}^{n} f(\tilde{X}_i) - \mathbb{E}\, f(X) \right) \right) \\
&\qquad \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \left(f(X_i) - f(\tilde{X}_i) \right) \leq \frac{2B}{n}.
\end{aligned}$$

Similarly, we have

$$G(\tilde{X}_1, \ldots, \tilde{X}_n) - G(X_1, \ldots, X_n) \geq -\frac{2B}{n}.$$

Therefore, the variation satisfies

$$L_i := \sup_{X, \tilde{X}} |G(X_1, \ldots, X_n) - G(\tilde{X}_1, \ldots, \tilde{X}_n)| \leq 2B/n,$$

where $X = (X_1, \ldots, X_n)$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ are different for only the i-th component.

Therefore, $\sigma^2 = \frac{1}{4} \sum_{i=1}^{n} L_i^2 \leq \frac{B^2}{n}$. By McDiarmid's inequality,

$$\mathbb{P}\{|G(X_1, \ldots, X_n) - \mathbb{E}\, G| \geq t\} \leq 2e^{-\frac{nt^2}{2B^2}}.$$

Let the failure probability $2e^{-\frac{nt^2}{2B^2}} = \delta$, which leads to $t = \sqrt{\frac{2B^2 \log(2/\delta)}{n}}$. Restating the above inequality gives the bound (13).

Analogously, we can apply McDiarmid's inequality to the Rademacher complexity $Q\left(x_1, \ldots, x_n\right) = \mathbb{E}_\xi \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \xi_i f\left(x_i\right) \right]$, which leads to the sample-dependent bound (14). □

Examples.

- Let $\mathcal{F} = \{f\}$. Then,

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_\xi [\frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i)] = 0.$$

- Two functions. Let $\mathcal{F} = \{f_{-1}, f_1\}$ where $f_{-1} \equiv -1$ and $f_1 \equiv 1$.

$$\sqrt{n} \widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_\xi \sup_{f \in \{-1, +1\}} f \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i = \mathbb{E}_\xi \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \right| \to \mathbb{E}_{Z \sim \mathcal{N}(0,1)} |Z| = \sqrt{\frac{2}{\pi}}.$$

  Hence, when n is sufficiently large,

$$\mathrm{Rad}_n(\mathcal{F}) \sim \sqrt{\frac{2}{n\pi}}.$$

Remark: This implies that it is impossible to obtain a rate faster than $O(1/\sqrt{n})$ using Rademacher complexity since it saturates even for learning/distinguishing two constant functions. This is bad news!

Lemma 5.4 (Massart's lemma)**.** *Assume that* $\sup_{x\in\mathcal{X},f\in\mathcal{F}}|f(x)|\le B$ *and* $\mathcal{F}$ *is finite. Then,*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F})\le B\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

Proof. Let $Z_f=\sum_{i=1}^n\xi_i f(x_i)$. Then,

$$\log\mathbb{E}[e^{\lambda Z_f}]=\log\left(\prod_{i=1}^n\mathbb{E}[e^{\lambda\xi_i f(x_i)}]\right)\le\sum_{i=1}^n\log\mathbb{E}\,e^{\lambda\xi_i f(X_i)}\overset{(i)}{\le}\sum_{i=1}^n\lambda^2\frac{(B-(-B))^2}{8}=\frac{nB^2}{2}\lambda^2,$$

where (i) follows from the Hoeffding's lemma, which provides an upper bound of the log-moment generating functions of a bounded random variable. Hence, Z_f is sub-Gaussian with the variance proxy $\sigma^2=nB^2$. Using the maximal inequality, we have

$$\widehat{\mathrm{Rad}}_n(\mathcal{F})=\frac{1}{n}\mathbb{E}_\xi[\sup_{f\in\mathcal{F}}Z_f]\le\frac{1}{n}\cdot\sqrt{n}B\sqrt{2\log|\mathcal{F}|}=B\sqrt{\frac{2\log|\mathcal{F}|}{n}}. \tag{15}$$

□

Applying Massart's lemma to bound the generalization gap recovers Lemma 3.1.

Linear functions. Let $\mathcal{F}=\{w^\top x:\|w\|_p\le 1\}$. Let q be the conjugate of p, i.e., $1/q+1/p=1$. Then,

$$\widehat{\mathrm{Rad}}_n(\mathcal{F})=\mathbb{E}_\xi\sup_{\|w\|_p\le1}\frac{1}{n}\sum_{i=1}^n\xi_i w^\top X_i=\mathbb{E}_\xi\sup_{\|w\|_p\le1}w^\top\left(\frac{1}{n}\sum_{i=1}^n\xi_i X_i\right)=\mathbb{E}_\xi\|\frac{1}{n}\sum_{i=1}^n\xi_i X_i\|_q. \tag{16}$$

Lemma 5.5. *Assume that* $\|x_i\|_q\le 1$ *for all* $i\in[n]$*. Then,*

- *If* $p=2$*, then*
$$\widehat{\mathrm{Rad}}_n(\mathcal{F})\le\sqrt{\frac{1}{n}}.$$
- *If* $p=1$*, then,*
$$\widehat{\mathrm{Rad}}_n(\mathcal{F})\le\sqrt{\frac{2\log(2d)}{n}}.$$

Proof. For the case where $p=2$,

$$\begin{aligned}\widehat{\mathrm{Rad}}_n(\mathcal{F})&\le\mathbb{E}_\xi\left\|\frac{1}{n}\sum_{i=1}^n\xi_i x_i\right\|_2\le\sqrt{\mathbb{E}_\xi\left\|\frac{1}{n}\sum_{i=1}^n\xi_i x_i\right\|_2^2}\\&=\sqrt{\frac{1}{n^2}\sum_{i,j=1}^n x_i x_j\,\mathbb{E}[\xi_i\xi_j]}=\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}\le\sqrt{\frac{1}{n}}.\end{aligned}$$

The case of $p=1$ is left as homework. □

We have shown the Rademacher complexity of linear functions. To obtain the estimates of more general classes, we need the following results.

Lemma 5.6 (Rademacher calculus)**.** *The Rademacher complexity has the following properties.*

- $\mathrm{Rad}_n(\lambda\mathcal{F}) = |\lambda| \mathrm{Rad}_n(\mathcal{F})$.
- $\mathrm{Rad}_n(\mathcal{F} + f_0) = \mathrm{Rad}_n(\mathcal{F})$.
- *Let Conv*$(\mathcal{F})$ *denote the convex hull of* $\mathcal{F}$ *defined by*

$$\mathit{Conv}(\mathcal{F}) = \Big\{ \sum_{j=1}^m a_j f_j : \alpha_j \geq 0, \sum_{j=1}^m a_j = 1, f_1, \ldots, f_m \in \mathcal{F}, m \in \mathbb{N}_+ \Big\}.$$

Then, we have $\mathrm{Rad}_n(\mathit{Conv}(\mathcal{F})) = \mathrm{Rad}_n(\mathcal{F})$.

Proof. Here, we only prove the third result. By definition,

$$\begin{aligned}
n\widehat{\mathrm{Rad}}_n(\mathrm{Conv}(\mathcal{F})) &= \mathbb{E} \sup_{f_j \in \mathcal{F}, \|\alpha\|_1 = 1} \sum_{i=1}^n \xi_i \sum_{j=1}^m a_j f_j(X_i) \\
&= \mathbb{E} \sup_{f_j \in \mathcal{F}, \|\alpha\|_1 = 1} \sum_{j=1}^m a_j \sum_{i=1}^n \xi_i f_j(X_i) \\
&= \mathbb{E} \sup_{f_j \in \mathcal{F}} \max_j \sum_{i=1}^n \xi_i f_j(X_i) \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \xi_i f(X_i) = n\widehat{\mathrm{Rad}}_n(\mathcal{F}).
\end{aligned}$$

□

The third property suggests that convex combinations do not change the Rademacher complexity.

Lemma 5.7 (Ledoux & Talagrand 2011, Contraction lemma)**.** *Let* $\varphi_i : \mathbb{R} \mapsto \mathbb{R}$ *with* $i = 1, \ldots, n$ *be* β*-Lipschitz continuous. Then,*

$$\frac{1}{n} \mathbb{E}_\xi \sup_{f \in \mathcal{F}} \sum_{i=1}^n \xi_i \varphi_i \circ f(x_i) \leq \beta \widehat{\mathrm{Rad}}_n(\mathcal{F}).$$

Proof. WLOG, assume $\beta = 1$. Let $\hat{\xi} = (\xi_1, \ldots, \xi_n)$ and $Z_k(f) = \sum_{i=1}^k \xi_i \varphi_i \circ f(x_i)$. Then,

$$\begin{aligned}
\mathbb{E}_{\xi_n} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \xi_i \varphi_i \circ f(x_i) &= \frac{1}{2} \left[\sup_{f \in \mathcal{F}} (Z_{n-1}(f) + \varphi_n \circ f(x_n)) + \sup_{f \in \mathcal{F}} (Z_{n-1}(f) - \varphi_n \circ f(x_n)) \right] \\
&= \frac{1}{2} \sup_{f, f' \in \mathcal{F}} \Big(Z_{n-1}(f) + Z_{n-1}(f') + \varphi_n \circ f(x_n) - \varphi_n \circ f'(x_n) \Big) \\
&\leq \frac{1}{2} \sup_{f, f' \in \mathcal{F}} \Big(Z_{n-1}(f) + Z_{n-1}(f') + |f(x_n) - f'(x_n)| \Big)
\end{aligned}$$

$$= \frac{1}{2} \sup_{f,f' \in \mathcal{F}} \Big(Z_{n-1}(f) + Z_{n-1}(f') + (f(x_n) - f'(x_n)) \Big) \quad \text{(Use the symmetry)}$$
$$= \frac{1}{2} \left[\sup_{f \in \mathcal{F}} (Z_{n-1}(f) + f(x_n)) + \sup_{f \in \mathcal{F}} (Z_{n-1}(f) - f(x_n)) \right]$$
$$= \mathbb{E}_{\xi_n} \sup_{f \in \mathcal{F}} (Z_{n-1}(f) + \xi_n f(x_n)).$$

Hence, by induction, we have

$$\begin{aligned}
\mathbb{E}_{\hat{\xi}}[\sup_{f \in \mathcal{F}} Z_n(f)] &\leq \mathbb{E}_{\hat{\xi}} \sup_{f \in \mathcal{F}} (Z_{n-1}(f) + \xi_n f(x_n)) \\
&\leq \mathbb{E}_{\hat{\xi}} \sup_{f \in \mathcal{F}} (Z_{n-2}(f) + \xi_{n-1} f(x_{n-1}) + \xi_n f(x_n)) \\
&\leq \mathbb{E}_{\hat{\xi}} \sup_{f \in \mathcal{F}} (\xi_1 f(x_1) + \cdots + \xi_n f(x_n)) \\
&= n\widehat{\mathrm{Rad}}_n(\mathcal{F}). \qquad (17)
\end{aligned}$$

□

Corollary 5.8. *Given a function class $\mathcal{F}$ and $\varphi : \mathbb{R} \mapsto \mathbb{R}$, let $\varphi \circ \mathcal{F} = \{\varphi \circ f : f \in \mathcal{F}\}$. Then,*

$$\mathrm{Rad}_n(\varphi \circ \mathcal{F}) \leq \mathit{Lip}(\varphi)\, \mathrm{Rad}_n(\mathcal{F}).$$

Rademacher complexity of neural networks. With the preceding results in hand, we can now directly derive a quite surprising result: the Rademacher complexity of two-layer neural networks. Specifically, let us examine the following steps that demonstrate how to construct a neural network from linear functions:

$$\begin{aligned}
\mathcal{L} &:= \{x \mapsto w^\top x : w \in \mathbb{S}^{d-1}\}, \\
\mathcal{L}_\sigma &:= \{x \mapsto \sigma(w^\top x) : w \in \mathbb{S}^{d-1}\}, \\
\mathcal{N}_\sigma &:= \left\{ x \mapsto \sum_{j=1}^{m} a_j \sigma(w_j^\top x) : m \in \mathbb{N},\, w_j \in \mathbb{S}^{d-1},\, a_j \geq 0,\, \sum_{j=1}^{m} a_j = 1 \right\}.
\end{aligned}$$

Here:

- $\mathcal{L}$ represents the set of linear functions with weights constrained to the unit sphere in $\mathbb{R}^d$.
- $\mathcal{L}_\sigma$ is formed by applying a nonlinear activation function σ to the linear functions in $\mathcal{L}$.
- $\mathcal{N}_\sigma$ is the convex hull of $\mathcal{L}_\sigma$, corresponding to functions expressed by two-layer neural networks with non-negative outer coefficients.

It is evident that $\mathcal{N}_\sigma$ describes the functions represented by two-layer neural networks where the outer coefficients are non-negative. Then, by applying Lemmas 5.5, 5.6, and 5.7, we obtain:

$$\mathcal{R}_n(\mathcal{N}_\sigma) = \mathcal{R}_n(\mathcal{L}_\sigma) \leq \mathrm{Lip}(\sigma) \cdot \mathcal{R}_n(\mathcal{L}) \leq \frac{\mathrm{Lip}(\sigma)}{\sqrt{n}}.$$

A particularly interesting observation is that the Rademacher complexity does not depend on the network width. This suggests that, with appropriate norm constraints, the complexity of over-parameterized models remains well-controlled. Thus, over-parameterization should not pose a significant problem.

Next, we provide a general derivation that applies to neural networks with arbitrary (i.e., possibly negative) outer coefficients. Consider two-layer neural networks. Suppose the activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is σ_{Lip}-Lipschitz continuous. Let

$$\mathcal{F}_m = \left\{ f_m(x;\theta) = \sum_{j=1}^{m} a_j \sigma(w_j^\top x) : \sum_j |a_j| \le A, \|w_j\|_2 \le B \right\}.$$

be the collection of two-layer neural networks $f_m(\cdot;\theta)$.

Lemma 5.9. *Suppose $\|x_i\|_2 \le 1$ for $i = 1, \ldots, n$. Then, we have*

$$\widehat{\text{Rad}}_n(\mathcal{F}_m) \le \frac{2\sigma_{\text{Lip}} AB}{\sqrt{n}}.$$

The above lemma implies that Rademacher complexity only depends on the parameter norm, independent of the network width. This implies that the capacity of over-parameterized networks can be well-controlled by enforcing a constraint on an appropriate parameter norm. It is worth noting that for different networks, we may need to identify the appropriate norm of parameters.

Proof.

$$\begin{aligned}
\widehat{\text{Rad}}_n(\mathcal{F}_m) &= \frac{1}{n}\mathbb{E}_\xi \sup_{f\in\mathcal{F}_m} \sum_{i=1}^{n} f(x_i)\xi_i \\
&= \frac{1}{n}\mathbb{E}_\xi \sup_{\theta\in\Theta} \sum_{i=1}^{n} \xi_i \sum_{j=1}^{m} a_j \sigma(w_j^\top x_i) \\
&= \frac{1}{n}\mathbb{E}_\xi \sup_{\theta\in\Theta} \sum_{j=1}^{m} a_j \sum_{i=1}^{n} \xi_i a_j \sigma(w_j^\top x_i) \\
&\le \frac{1}{n}\mathbb{E}_\xi \sup_{\theta\in\Theta} \sum_{j=1}^{m} |a_j| \left| \sup_{\|w\|\le B} \sum_{i=1}^{n} \xi_i \sigma(w^\top x_i) \right| \\
&\overset{(i)}{\le} A\frac{1}{n}\mathbb{E}_\xi \sup_{\|w\|\le B} \left| \sum_{i=1}^{n} \xi_i \sigma(w^\top x_i) \right| \\
&= A\frac{1}{n}\mathbb{E}_\xi \left(\sup_{\|w\|\le B} \sum_{i=1}^{n} \xi_i \sigma(w^\top x_i) \right) + A\frac{1}{n}\mathbb{E}_\xi \left(- \sup_{\|w\|\le B} \sum_{i=1}^{n} \xi_i \sigma(w^\top x_i) \right) \\
&\overset{(ii)}{\le} 2A\frac{1}{n}\mathbb{E}_\xi \left(\sup_{\|w\|\le B} \sum_{i=1}^{n} \xi_i \sigma(w^\top x_i) \right) \\
&\overset{iii}{\le} 2A\sigma_{\text{Lip}}\frac{1}{n}\mathbb{E}_\xi \left(\sup_{\|w\|\le B} \sum_{i=1}^{n} \xi_i w^\top x_i \right)
\end{aligned}$$

$$\overset{(iiii)}{\leq} \frac{2\sigma_{\text{Lip}}AB}{\sqrt{n}},$$

where (i) is due to $\sum_{j=1}^{m}|a_j| \leq A$; (ii) use the symmetry of ξ_i; (iii)follows from the contraction property (Lemma 5.7); $(iiii)$ follows from Lemma 5.5. □

6 Bounding Rademacher complexity using covering number

Consider the function space $(\mathcal{F}, L^2(\mathbb{P}_n))$, where $\mathcal{F}$ is the hypothesis class and $L^2(\mathbb{P}_n)$ is defined by

$$\|f - f'\|_{L^2(\mathbb{P}_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f'(x_i))^2},$$

where $x_1, \ldots, x_n$ denote the finite training samples. Since only the n samples are available, we can really think of these functions as a n-dimensional vector:

$$\hat{f} = (f(x_1), f(x_2), \ldots, f(x_n))^\top \in \mathbb{R}^n.$$

Obviously, we cannot distinguish functions using information beyond these n-dimensional vectors.

Example 1. Let $\mathcal{F} = \{f : \mathbb{R} \mapsto [0,1] : f \text{ is non-decreasing}\}$. Then, $\mathcal{N}(\mathcal{F}, L_2(\mathbb{P}_n), \varepsilon) = n^{1/\varepsilon}$.

This function class is important for two reasons: 1) it encompasses cumulative distribution functions; 2) it is associated with monotone single-index models, which involve learning functions of the form $x \mapsto \sigma(w^\top x)$ where σ is an unknown monotone function. These models are prevalent in econometrics, biostatistics, and machine learning.

Proof. WLOG, assume $-\infty = x_0 < x_1 \leq x_2 \leq \cdots \leq x_n \leq x_{n+1} = 1$. For any $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, define a piecewise constant function

$$f_y(x) = y_i \quad \text{for } x \in [x_i, x_{i+1}),\ i = 1, 2, \ldots, n.$$

For any $\varepsilon \in (0,1)$, let $Y_\varepsilon = (0, \varepsilon, 2\varepsilon, 3\varepsilon, \ldots, 1-\varepsilon)$. Then, $|Y_\varepsilon| \leq 1/\varepsilon$. Define the following non-decreasing set:

$$S_\varepsilon := \{y \in \mathbb{R}^n : y_i \in Y_\varepsilon \text{ and } y_1 \leq y_1 \leq \cdots \leq y_n\}.$$

Let $\mathcal{F}_\varepsilon = \{f_y : y \in S_\varepsilon\}$. Obviously, $\mathcal{F}_\varepsilon \subset \mathcal{F}$. Moreover, for any $f \in \mathcal{F}$, there exists $y \in S_\varepsilon$ such that

$$\|f - f_y\|^2_{L_2(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \leq \varepsilon^2.$$

Hence, $\mathcal{F}_\varepsilon$ is an ε-cover of $\mathcal{F}$ and $|\mathcal{F}_\varepsilon| = |S_\varepsilon|$. What remains is to count the cardinality of $|S_\varepsilon|$. Let $y_0 = 0, y_{n+1} = 1$ and $\Delta_i = (y_i - y_{i-1})/\varepsilon$. Then, $\{\Delta_i\}_{i=1}^{n+1}$ must be non-negative integers and satisfy

$$\Delta_1 + \Delta_2 + \ldots \Delta_{n+1} = \frac{1}{\varepsilon}.$$

Hence, $|S_\varepsilon|$ is equal to the number of solutions to the above equation:

$$|S_\varepsilon| = \binom{n+\frac{1}{\varepsilon}}{n} = \frac{(n+\frac{1}{\varepsilon})(n+\frac{1}{\varepsilon}-1)\cdots(n+1)}{(\frac{1}{\varepsilon})(\frac{1}{\varepsilon}-1)\cdots 1} \le n^{\frac{1}{\varepsilon}}.$$

□

In the following, we show that the Rademacher complexity can be bounded using the metric entropy. To simplify notation, we use $\|\cdot\|$ and $\langle,\rangle$ to denote $L^2(\mathbb{P}_n)$ norm and the induced inner product: $\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i)$. Then,

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}}\langle \xi, f\rangle.$$

Proposition 6.1 (One-resolution discretization)**.** *Suppose* $\sup_{x\in\mathcal{X},f\in\mathcal{F}}|f(x)| \le B$*. Then,*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) \le \inf_\varepsilon \left(\varepsilon + B\sqrt{\frac{2\log \mathcal{N}(\mathcal{F}, L_2(\mathbb{P}_n), \varepsilon)}{n}}\right).$$

The above bound is similar to Theorem 4.4. The difference is that the above bound is determined by the $L^2(\mathbb{P}_n)$ covering number, while Theorem 4.4 relies on the L^∞ covering number. Technically speaking, this improvement is obtained by removing the $\mathbb{E}\, f(X)$ term with symmetrization.

Proof. Let $\mathcal{F}_\varepsilon$ be an ε-cover of $\mathcal{F}$ with respect to the metric $L^2(\mathbb{P}_n)$. For any $f \in \mathcal{F}$, let $\pi(f) \in \mathcal{F}_\varepsilon$ such that $\|f - \pi(f)\| \le \varepsilon$. Then,

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\langle \xi, f\rangle &= \mathbb{E}\sup_{f\in\mathcal{F}}\Big[\langle \xi, f-\pi(f)\rangle + \langle \xi, \pi(f)\rangle\Big] \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\langle \xi, f-\pi(f)\rangle + \mathbb{E}\sup_{f\in\mathcal{F}}\langle \xi, \pi(f)\rangle \\
&\le \mathbb{E}\,\|\xi\|\|f-\pi(f)\| + \mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}\langle \xi, f\rangle \\
&\le \varepsilon\sqrt{\frac{\mathbb{E}\,\|\xi\|_2^2}{n}} + \widehat{\mathrm{Rad}}_n(\mathcal{F}_\varepsilon) \qquad \text{(Jesson's inequality)} \\
&\le \varepsilon + B\sqrt{\frac{2\log|\mathcal{F}_\varepsilon|}{n}}, \qquad \text{(Massart's lemma).}
\end{aligned}$$

Using the definition of covering number and optimizing over ε, we complete the proof. □

For the non-decreasing functions considered previously, we have

$$\mathrm{Rad}_n(\mathcal{F}) \le \inf\left(\varepsilon + \sqrt{\frac{2\log n}{\varepsilon n}}\right) = C\left(\frac{\log n}{n}\right)^{1/3}. \tag{18}$$

This rate is slower than the expected $O(1/\sqrt{n})$. Is it because non-decreasing functions are complex? No! It is actually just an artifact caused by the proof technique.

In many cases, the one-resolution discretization may give us sub-optimal bounds of the generalization gap. To fix this problem, we need a sophisticated analysis of all the resolutions. This is typically done by using a *chaining* approach introduced by Dudley.

Theorem 6.2 (Dudley's integral inequality)**.** *Let $D = \sup_{f,f' \in \mathcal{F}} \|f - f'\|_{L^2(\mathbb{P}_n)}$ be the diameter of $\mathcal{F}$. Then,*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) \leq 12 \inf_{\alpha \in [0,D]} \left(\alpha + \int_\alpha^D \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(\mathbb{P}_n), \varepsilon)}{n}} \, \mathrm{d}\varepsilon \right).$$

Then, for the for non-decreasing functions, we have

$$\mathrm{Rad}_n(\mathcal{F}) \lesssim \int_0^2 \sqrt{\frac{\log n}{n\varepsilon}} \, \mathrm{d}\varepsilon \lesssim \sqrt{\frac{\log n}{n}}.$$

Figure 1 visualizes the difference between the upper bound given in Proposition 6.1 and the one in Theorem 6.2. Clearly, the latter is smaller.

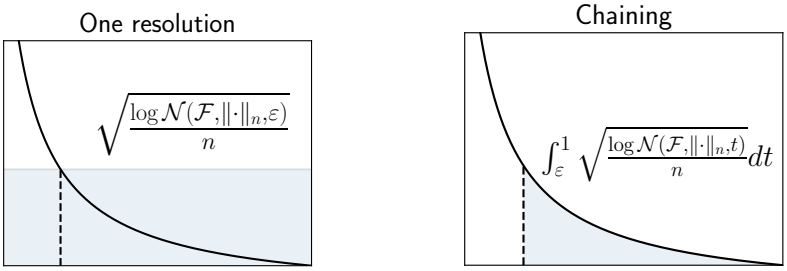

Figure 1: (Left) The result of one-resolution analysis; (Right) The result of chaining with all resolutions. In this case, the diameter $D = 1$. The comparison of two figures provides a visual illustration of how the chaining bound is tigher than the one-resolution bound.

Proof. Let $\varepsilon_j = 2^{-j} D$ be the dyadic scale and $\mathcal{F}_j$ be an ε_j-cover of $\mathcal{F}$. Given $f \in \mathcal{F}$, let $f_j \in \mathcal{F}_j$ such that $\|f_j - f\| \leq \varepsilon_j$. Consider the decomposition

$$f = f - f_m + \sum_{j=1}^m (f_j - f_{j-1}), \tag{19}$$

where $f_0 = 0$. Notice that

- $\|f - f_m\| \leq \varepsilon_m$.
- $\|f_j - f_{j-1}\| \leq \|f_j - f\| + \|f - f_{j-1}\| \leq \varepsilon_j + \varepsilon_{j-1} \leq 3\varepsilon_j$.

Then,

$$\begin{aligned}
\widehat{\mathrm{Rad}}_n(\mathcal{F}) &= \mathbb{E} \sup_{f \in \mathcal{F}} \langle \xi, f \rangle \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \left(\langle \xi, f - f_m \rangle + \sum_{j=1}^m \langle \xi, f_j - f_{j-1} \rangle \right) \\
&\leq \varepsilon_m + \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{j=1}^m \langle \xi, f_j - f_{j-1} \rangle
\end{aligned}$$

$$
\begin{aligned}
&\le \varepsilon_m + \sum_{j=1}^{m} \mathbb{E} \sup_{f\in\mathcal{F}} \langle \xi, f_j - f_{j-1}\rangle \\
&= \varepsilon_m + \sum_{j=1}^{m} \mathbb{E} \sup_{f_j\in\mathcal{F}_j, f_{j-1}\in\mathcal{F}_{j-1}} \langle \xi, f_j - f_{j-1}\rangle \\
&= \varepsilon_m + \sum_{j=1}^{m} \widehat{\mathrm{Rad}}_n(\mathcal{F}_j \cup \mathcal{F}_{j-1}).
\end{aligned}
$$

Using the Massart lemma and the fact that $\sup_{f\in\mathcal{F}_j, f'\in\mathcal{F}_{j-1}} \|f_j - f_{j-1}\| \le 3\varepsilon_j$,

$$
\begin{aligned}
\widehat{\mathrm{Rad}}_n(\mathcal{F}) &\le \varepsilon_m + \sum_{j=1}^{m} 3\varepsilon_j \sqrt{\frac{2\log(|\mathcal{F}_j||\mathcal{F}_{j-1}|)}{n}} \\
&\le \varepsilon_m + \sum_{j=1}^{m} 6\varepsilon_j \sqrt{\frac{\log|\mathcal{F}_j|}{n}} \\
&= \varepsilon_m + \sum_{j=1}^{m} 12(\varepsilon_j - \varepsilon_{j+1}) \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(\mathbb{P}_n), \varepsilon_j)}{n}}.
\end{aligned}
$$

Taking $m \to \infty$, we obtain

$$
\widehat{\mathrm{Rad}}_n(\mathcal{F}) \le 12 \int_0^D \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(\mathbb{P}_n), t)}{n}}\, \mathrm{d}t.
$$

□

Similarly, we can obtain that

$$
\widehat{\mathrm{Rad}}_n(\mathcal{F}) \lesssim \inf_{\alpha>0} \left(\alpha + \int_\alpha^D \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(\mathbb{P}_n), t)}{n}}\, \mathrm{d}t \right).
$$

The key ingredient of proceeding analysis is the multi-resolution decomposition (19). The technical reason why chaining provides a better estimate is as follows. In the one-resolution discretization, we apply Massart's lemma to functions whose range in $[-1, 1]$, whereas in chaining, we apply Massart's lemma to functions whose range has size $O(\varepsilon_j)$.

Remark 6.3. *Can we prove the same result by considering uniform resolutions $\{\epsilon_j = jD/m\}$, where m is chosen such that $D/m \le \alpha$?*

Remark 6.4. *Metric entropy is often more intuitive than Rademacher complexity, as it is fundamentally based on discretization and the application of Massart's lemma. Additionally, in many cases, metric entropy is more convenient to estimate. However, Rademacher complexity can sometimes provide sharper upper bounds when it can be directly estimated, without relying on covering number estimates and the subsequent application of the Dudley integral bound.*

References

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.