Theoretical Deep Learning

# Lecture 3: Uniform bounds and empirical processes

July 27, 2021

*Lecturer: Lei Wu* *Scribe: Lei Wu*

## 1 Uniform bounds of generalization gap

Let $\mathcal{H}$ be the hypothesis class. Consider the estimator:

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

This estimator guarantees the smallness of the empirical risk. But the question is: How small is the true error $\mathcal{R}(\hat{h}_n)$? This is equivalent to control the generalization gap:

$$\mathcal{R}(\hat{h}_n) - \hat{\mathcal{R}}_n(\hat{h}_n). \tag{1.1}$$

Unfortunately, concentration inequalities cannot be applied directly since $\hat{h}_n$ depends on the training set. To deal with this dependence, we can consider the uniform bound

$$|\mathcal{R}(\hat{h}_n) - \hat{\mathcal{R}}_n(\hat{h}_n)| \leq \sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}_n(h)|. \tag{1.2}$$

Obviously, when the hypothesis space $\mathcal{H}$ is sufficiently "small", e.g., the extreme case: $\mathcal{H} = \{h\}$, it is expected that

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

Some natural questions go as follows.

- What kind of $\mathcal{H}$ can guarantee the smallness of uniform bound?
- What is the rate? Do we still have $O(1/\sqrt{n})$?

Let us first look at a simple example: finite hypothesis class.

**Lemma 1.1.** *Assume $|\mathcal{H}| < \infty$ and $\sup_{y,y'} |\ell(y,y')| \leq 1$. For any $\delta \in (0,1)$, with probability $1-\delta$ over the random sampling of training set $S$, we have*

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

*Proof.* Let $Z(h,X) = \ell(h(X), h^*(X))$. Taking the union bound gives us

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} |\frac{1}{n}\sum_{i=1}^n Z(h,X_i) - \mathbb{E}[Z(h,X)]| \geq t\right\} \leq \sum_{j=1}^m \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z(h_j,X_i) - \mathbb{E}[Z(h_j,X)]\right| \geq t\right\} \tag{1.3}$$

$$\leq m2e^{\frac{-2nt^2}{2^2}} = 2me^{\frac{-nt^2}{2}}. \tag{1.4}$$

Let the failure probability $2me^{\frac{-nt^2}{2}} = \delta$, which leads to $t = \sqrt{\frac{2\ln(2m/\delta)}{n}}$.

□

The upper bound only depends on $|\mathcal{H}|$ logarithmically. Hence, even when the hypothesis class has exponentially many functions, the generalization gap can be still well controlled.

**Definition 1.2** (Empirical process)**.** Let $\mathcal{F}$ be a class of real-valued functions $f : \Omega \mapsto \mathbb{R}$ where $(\Omega, \Sigma, \mu)$ is a probability space. Let $X \sim \mu$ and $X_1, \dots, X_n$ be independent copies of $X$. Then, the random process $(X_f)_{f\in\mathcal{F}}$ defined by

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

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

In our case, $f(X) = \ell(h(X), h^*(X))$. Our task is to bound the suprema:

$$\sup_{f\in\mathcal{F}} |X_f|.$$

Note that the above quantity can viewed a "weak" distance between $\mu$ and the empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta(\cdot - x_i)$ with the test functions given by $\mathcal{F}$:

$$d_{\mathcal{F}}(\hat{\mu}_n, \mu) := \sup_{f\in\mathcal{F}} |\, \mathbb{E}_{\hat{\mu}_n} f - \mathbb{E}_{\mu} f|.$$

# 2 Rademacher complexity

**Lemma 2.1** (Symmetrization of empirical processes)**.**

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

*where $\xi_1, \dots, \xi_n$ are i.i.d. Rademacher random variable:* $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = \frac{1}{2}$

*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,

$$\mathbb{E} \sup_{f\in\mathcal{F}} [\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)] = \mathbb{E}_{X_i} \sup_{f\in\mathcal{F}} \mathbb{E}_{X_i'} [\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - f(X_i'))] \tag{2.1}$$

$$\le \mathbb{E}_{X_i, X_i'} \sup_{f\in\mathcal{F}} [\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - f(X_i'))] \tag{2.2}$$

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

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

$$= 2\,\mathbb{E}_{X_i,\xi} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \xi_i f(X_i)$$

□

**Definition 2.2** (Rademacher complexity)**.** The empirical Rademacher complexity of a function class $\mathcal{F}$ on finite samples 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

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

The symmetrization lemma 2.1 implies that

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)\right] \le 2\,\mathrm{Rad}_n(\mathcal{F}). \tag{2.3}$$

**Theorem 2.3.** *Assume that $0 \le f \le B$ 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*

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

*and the sample-dependent version:*

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E}f(X)\right| \le 2\widehat{\mathrm{Rad}}_n(\mathcal{F}) + 3B\sqrt{\frac{\log(4/\delta)}{n}}.$$

*Proof.* Let

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

and note that

$$\sup_{\alpha} g\left(x_1, \ldots, x_{i-1}, \alpha, x_{i+1}, \ldots, x_n\right) - \inf_{\alpha} g\left(x_1, \ldots, x_{i-1}, \alpha, x_{i+1}, \ldots, x_n\right) \le \frac{B}{n}.$$

By McDiarmid's inequality,

$$\mathbb{P}\{|g(X_1, \ldots, X_n) - \mathbb{E}\, g| \ge t\} \le 2e^{-\frac{2nt^2}{B^2}}.$$

Let the failure probability $2e^{-\frac{2nt^2}{B^2}} = \delta$, which leads to $t = \sqrt{\frac{2B\log(2/\delta)}{n}}$. This proves the first statement. Analogously, using again the McDiarmid's inequality to $g'\left(x_1, \ldots, x_n\right) = \mathbb{E}_\xi \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} \xi_i f\left(x_i\right)\right]$ leads to the sample-dependent one. □

- 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 $f_1 \equiv 1$.

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

  Hence, when $n$ is sufficiently large,

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

**Lemma 2.4** (Massart's lemma)**.** *Assume that* $\sup_{x\in\mathcal{X}, f\in\mathcal{F}} |f| \leq B$ *and* $\mathcal{F}$ *is finite. Then,*

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

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

$$\mathbb{E}[e^{\lambda Z_f}] = \prod_{i=1}^n \mathbb{E}[e^{\lambda \xi_i f(x_i)}] \leq \prod_{i=1}^n e^{\lambda^2 \frac{(B-(-B))^2}{8}} = e^{\frac{\lambda^2 n B^2}{2}}.$$

Hence, $Z_f$ is sub-Gaussian with the variance proxy $\sigma^2 = \sqrt{n}B$. Using the maximal inequality, we have

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\xi[\sup_{f\in\mathcal{F}} Z_f] \leq \frac{1}{n}\cdot\sqrt{n}B\sqrt{2\log|\mathcal{F}|} = B\sqrt{\frac{2\log|\mathcal{F}|}{n}}. \tag{2.4}$$

□

Applying Massart's lemma to bound the generalization gap recovers Lemma 1.1.

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

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_\xi \sup_{\|w\|_p\leq 1} \frac{1}{n}\sum_{i=1}^n \xi_i w^T X_i = \mathbb{E}_\xi \sup_{\|w\|_p \leq 1} w^T\left(\frac{1}{n}\sum_{i=1}^n \xi_i X_i\right) = \mathbb{E}_\xi \,\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\|_q. \tag{2.5}$$

**Lemma 2.5.** *Assume that* $\|x_i\|_q \leq 1$ *for all* $x_i \in S$. *Then,*

- *If* $p = 2$, *then*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) \leq \sqrt{\frac{1}{n}}.$$

- *If* $p = 1$, *then,*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(2d)}{n}}.$$

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

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

The case of $p = 1$ leaves to homework. □

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

**Lemma 2.6** (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*

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

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

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

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

□

**Lemma 2.7** (Ledoux & Talagrand 2011, Contraction lemma)**.** *Let* $\varphi_i : \mathbb{R} \mapsto \mathbb{R}$ *with* $i = 1, \dots, n$ *be* $\beta$*-Lispchitz continuous. Then,*

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

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

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

Hence, by induction, we have

$$\begin{aligned}
\mathbb{E}_{\hat{\xi}}[\sup_{f\in\mathcal{F}} Z_n(f)] &\leq \mathbb{E}_{\hat{\xi}} \sup_{f\in\mathcal{F}}(Z_{n-1}(f) + \xi_n f(x_n)) \\
&\leq \mathbb{E}_{\hat{\xi}} \sup_{f\in\mathcal{F}}(Z_{n-2}(f) + \xi_{n-1} f(x_{n-1}) + \xi_n f(x_n)) \\
&\leq \mathbb{E}_{\hat{\xi}} \sup_{f\in\mathcal{F}}(\xi_1 f(x_1) + \cdots + \xi_n f(x_n)) \\
&= n\widehat{\mathrm{Rad}}_n(\mathcal{F}). \qquad (2.6)
\end{aligned}$$

$\square$

**Corollary 2.8.** *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,*

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

# 3 Covering number and 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. Can we extend this observation to the case where $|\mathcal{F}| = \infty$. One possible approach is *discretization*. This means that we choose a finite subset $\mathcal{F}_\varepsilon \subset \mathcal{F}$ to "represent" $\mathcal{F}$.

**Definition 3.1.** Consider a metric space $(T, \rho)$.

- We say $T_\varepsilon \subset T$ is a $\varepsilon$-cover (also called $\varepsilon$-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 $\mathcal{N}(\varepsilon, T, \rho)$ is defined as the smallest cardinality of an $\varepsilon$-cover of $T$ with respect to $\rho$. The *metric entropy* of $T$ is defined by $\log \mathcal{N}(\varepsilon, T, \rho)$.

In the above definition, the metric space $(T, \rho)$ can be arbitrary. However, we will focus on the case of $(\mathcal{F}, L^2(\mathbb{P}_n))$, where $\mathcal{F}$ is the hypothesis class and $L^2(\mathbb{P}_n)$ is defined by

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

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

$$\hat{f} = (f(x_1), f(x_2), \dots, f(x_n))^T \in \mathbb{R}^n,$$

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, $\mathcal{N}(\varphi, \mathcal{F}, L_2(\mathbb{P}_n)) = n^{1/\varepsilon}$.

*Proof.* Given $\varepsilon \in (0,1)$, let $I_\varepsilon = (0, \varepsilon, 2\varepsilon, 3\varepsilon, \dots, 1)$. WLOG, assume $-\infty = x_0 < x_1 \leq x_2 \leq \cdots \leq x_n \leq x_{n+1} = 1$. Let $q(y;\varepsilon) = \operatorname{argmin}_{t \in I_\varepsilon} |t - y|$ for $y \in [0,1]$. For any $f \in \mathcal{F}$, define a piecewise approximation of $f$ as follows

$$\hat{f}_\varepsilon(x) = q(f(x_i);\varepsilon), \quad \text{for } x \in [x_i, x_{i+1}],\ i = 0, 1, \dots, n. \tag{3.1}$$

Using the monotonicity, we have $\|f - \hat{f}_\varepsilon\|_{L^2(\mathbb{P}_n)} \leq \varepsilon$. Then, let $\mathcal{F}_\varepsilon$ denote the set of all the piecewise constant functions defined above. By construction, $\mathcal{F}_\varepsilon$ is an $\varepsilon$-cover of $\mathcal{F}$. Moreover, $|\mathcal{F}_\varepsilon| \leq n^{1/\varepsilon}$. □

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 3.2** (One-step discretization)**.** *Suppose* $\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| \leq B$. *Then,*

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

*Proof.* Let $\mathcal{F}_\varepsilon$ be an $\varepsilon$-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 $\|f - \pi(f)\| \leq \varepsilon$. Then,

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

Using the definition of covering number and optimizing over $\varepsilon$, we complete the proof. □

For the non-decreasing functions considered previously, we have

$$\mathrm{Rad}_n(\mathcal{F}) \leq \inf \left( \varepsilon + \sqrt{\frac{2 \log n}{\varepsilon n}} \right) = C \left( \frac{\log n}{n} \right)^{1/3}. \tag{3.2}$$

This rate is slower than the expected $1/\sqrt{n}$. 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-step discretization may give us sub-optimal bounds of 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 3.3** (Dudley's integral inequality). *Assume* $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f - f'\|_{L^2(\mathbb{P}_n)} = D$ *be the diameter of* $\mathcal{F}$. *Then,*

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

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 3.2 and the one in Theorem 3.3. Clearly, the latter is smaller.

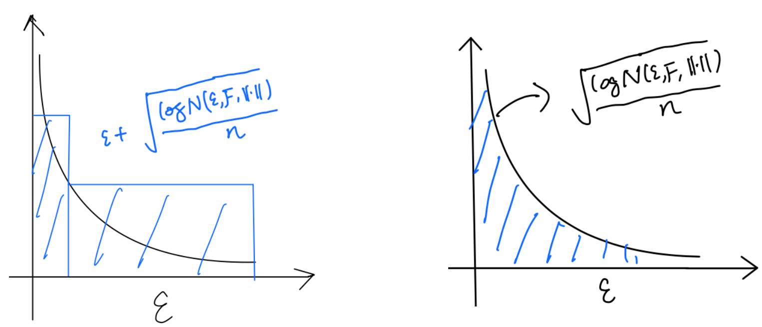


**Figure 1:** (Left) The result of one-resolution analysis; (Right) The result of chaining.

*Proof.* Let $D = \sup_{f,f' \in \mathcal{F}} \|f_1 - f_2\|$ be the diameter of $\mathcal{F}$. Let $\mathcal{F}_j$ be a $\varepsilon_j$-cover of $\mathcal{F}$ with $\varepsilon_j = 2^{-j} D$ be the dyadic scale. Let $f_j \in \mathcal{F}_j$ such that $\|f_j - f\| \leq \varepsilon_j$. Consider the decomposition

$$f = f - f_m + \sum_{j=1}^{m} (f_j - f_{j-1}), \tag{3.3}$$

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,

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

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

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

Taking $m\to\infty$, we obtain

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

□

The key ingredient of proceeding analysis is the multi-resolution decomposition (3.3). 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* 3.4. Metric entropy is actually a more intuitive complexity measure than Rademacher complexity. The essence is discretization and applying Massart's lemma. Moreover, metric entropy is sometimes more convenient to estimate.