**COMP 7070 Advanced Topics in AI and ML** **March 5, 2025**

# Lecture 7: Generalization Error

*Instructor: Yifan Chen* *Scribes: Bojian Zhu, Huiling Li* *Proof reader: Xiong Peng*

**Note:** *LaTeX template courtesy of UC Berkeley EECS department.*
**Disclaimer:** *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 7.1 Error Decomposition

1. Recall population risk and empirical risk. Given a distribution $P$ and a model function $f$, assume that a set of labeled data points is sampled from the distribution:

   - **Population Risk:**
   $$\mathcal{R}(f) = \mathbb{E}_{x,y\sim P}\, \ell(f(x), y);$$
   - **Empirical Risk:**
   $$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i).$$

2. In practice, assume that $\hat{f} \in \mathcal{F}$ and a reference $\bar{f} \in \mathcal{F}$[1].

3. Now, we start to perform error decomposition:

$$\mathcal{R}(\hat{f}) = \underbrace{\Big[\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\Big]}_{\textit{i. generalization}} + \underbrace{\Big[\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(\bar{f})\Big]}_{\textit{ii. optimization}} + \underbrace{\Big[\hat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f})\Big]}_{\textit{iii. concentration/generalization}} + \underbrace{\mathcal{R}(\bar{f})}_{\textit{iv. approximation error}} \tag{7.1}$$

   Four components are included in Eq. (7.1). Among them, the generalization term contributes to achieving good performance on both the training and testing sets, while the concentration term contributes to pushing the empirical risk closer to the population risk.

4. Three problems in deep learning theory (DLT):

   - **Representation:** related to term *iv*;
   - **Optimization:** related to term *ii*;
   - **Generalization:** related to terms *i* and *iii*.

## 7.2 Generalization

Consider an infinite function space $\mathcal{F}$; otherwise, we can use the union bound. Then, we can bound the space with *Rademacher Complexity.* The key spirit of Rademacher complexity, which focuses on a smaller proxy set, is similar to that of an $\varepsilon$-net.

[1] Note that $\bar{f}$ is not the optimal.

### 7.2.1 Un-normalized Rademacher Complexity

Given a collection of vectors $\mathcal{V}$, the un-normalized Rademacher complexity is defined as:

$$\mathrm{URad}(\mathcal{V}) := \mathbb{E}_\epsilon \sup_{a\in\mathcal{V}} \langle a, \epsilon\rangle. \tag{7.2}$$

### 7.2.2

Assume that

$$\mathcal{V} = \{(\ell(f(x_1), y_1), \ell(f(x_2), y_2), \ldots, \ell(f(x_n), y_n)) : f \in \mathcal{F}\}.$$

By applying Rademacher complexity to a dataset

$$\mathcal{S} = \{s_i = (x_i, y_i)\}_{i=1}^n,$$

then the Rademacher complexity of $\ell \circ \mathcal{F}_{|S}$ is given by:

$$\mathrm{Urad}(\ell \circ \mathcal{F}_{|S}) = \mathbb{E}_\epsilon \sup_{u\in\ell\circ\mathcal{F}_{|S}} \langle \epsilon, u\rangle = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^n \epsilon_i\, \ell(f(x_i), y_i)\right].$$

Note that $\mathrm{Urad}(\ell \circ \mathcal{F}_{|S})$ is a random variable, depending on $(x_i, y_i)$.

### 7.2.3

Here, we redefine the notation as $f(z_i) = \ell(f(x_i), y_i)$. Let $f(z) \in [a, b]$, $\forall f \in \mathcal{F}$, then with probability at least $1 - \delta$, we have:

$$\mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i) \le \sup_{f\in\mathcal{F}} \mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i) \le \mathbb{E}_{z_i}\left(\sup_{f\in\mathcal{F}} \mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i)\right) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

**Remark.** Based on the above summary, we can further have

$$\left|\sup_{f\in\mathcal{F}} \mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}_{z_i}\right| \le 2(b-a)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}.$$

**Lemma 7.1** *Given two functions $f$ and $g$, we have*

$$\sup_a \big(f(a) + g(a)\big) \le \sup_{a^*}\big(\sup_a f(a) + g(a^*)\big).$$

**Proof of Lemma 7.1**

*Proof.* Firstly, we know that for any $\varepsilon > 0$, there exists $a^*$ such that

$$f(a^*) + g(a^*) + \varepsilon \ge \sup_a f(a) + g(a).$$

Then, we know

$$\sup_a f(a) + g(a^*) \ge f(a^*) + g(a^*) \ge \sup_a f(a) + g(a) - \varepsilon.$$

Thus,

$$RHS = \sup_{a^*}\big(\sup_a f(a) + g(a^*)\big) \ge \sup_a f(a) + g(a).$$

**Lemma 7.2.** *Given two functions $f$ and $g$, we have*

$$-\sup_a \big(f(a)+g(a)\big) \leq \sup_{a^*}\big(-\sup_a f(a) - g(a^*)\big).$$

**Proof.**
*Step 1: Setup and Notation.*
Firstly, we introduce $\sup_{f\in\mathcal{F}} \mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i) \sim$ subG and it is a function of $z \sim z_n$ with bounded differences. Then, we have

$$\left|\sup_f \left(\mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i)\right) - \sup_{f'}\left(\mathbb{E}_z f'(z) - \frac{1}{n}\sum_{i=1}^n f'(z_i^{\setminus j})\right)\right|, \tag{7.3}$$

where we denote $z_i^{\setminus j} = \{z_1, z_2, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n\}$. For convenience, we reload the notations as follows:

$$\begin{aligned}
\mathbb{E}f &= \mathbb{E}_z f(z),\\
\mathbb{E}_n &= \mathbb{E}_{z_1\sim z_n},\\
\hat{\mathbb{E}}_n f &= \frac{1}{n}\sum_{i=1}^n f(z_i),\\
\mathbb{E}'_n &= \mathbb{E}_{z'_1\sim z'_n},\\
\hat{\mathbb{E}}'_n &= \frac{1}{n}\sum_{i=1}^n f(z'_i).
\end{aligned}$$

*Step 2: Reformulating the Key Quantity.*
Then, Eq. (7.3) can be further formulated as:

$$\begin{aligned}
&\left|\sup_f\Big(\mathbb{E}_z f(z) \ - \ \tfrac{1}{n}\sum_{i=1}^n f(z_i)\Big) \ - \ \sup_{f'}\Big(\mathbb{E}_z f'(z) \ - \ \tfrac{1}{n}\sum_{i=1}^n f'(z_i^{\setminus j})\Big)\right|\\
&= \ \left|\sup_f\big(\mathbb{E}f \ - \ \hat{\mathbb{E}}_n f\big) \ - \ \sup_{f'}\Big(\mathbb{E}f' \ - \ \hat{\mathbb{E}}_n f'\Big)\right| \\
&\leq \ \sup_{f''}\left|\sup_f\big(\mathbb{E}f \ - \ \hat{\mathbb{E}}_n f\big) \ - \ \sup_{f'}\big(\mathbb{E}f' \ - \ \hat{\mathbb{E}}_n f'\big) \ - \ \tfrac{1}{n}\big(f''(z_i) \ - \ f''(z'_i)\big)\right|.
\end{aligned} \tag{7.4}$$

*Step 3: Applying Lemma 7.2.*
With Lemma 7.2, we have

$$\begin{aligned}
\left|\sup_f\left(\mathbb{E}_f - \hat{\mathbb{E}}_n f\right)\right| &= \left|C - \sup_a\left(f(a)+g(a)\right)\right|\\
&\leq \sup_{a^*}\left|C - \sup_a f(a) - g(a^*)\right|.
\end{aligned} \tag{7.5}$$

*Step 4: Bounding the Dual Term.*
Similarly, we have

$$\begin{aligned}
&-C + \sup_a\left(f(a)+g(a)\right) \leq -C + \sup_{a^*}\left(\sup_a f(a) + g(a^*)\right)\\
&= \sup_{a^*}\left(-C + \sup_a f(a) + g(a^*)\right) \leq \sup_{a^*}\left|C - \sup_a f(a) - g(a^*)\right|.
\end{aligned}$$

*Step 5: Finalizing the Bound.*
Thus, we have

$$\sup_{f''}\left|\frac{1}{n}f''(z_i) - f''(z_i')\right| \le \frac{1}{n}(b-a).$$

Since $\sup_{f\in\mathcal{F}} \mathbb{E}f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)$ is sub-Gaussian with $\sigma^2 \le \frac{(b-a)^2}{4n}$, then

$$\mathbb{P}\Bigg(\sup_{f\in\mathcal{F}} \mathbb{E}f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}\Big(\sup_{f\in\mathcal{F}} \mathbb{E}f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\Big) > t\Bigg) \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big) = \delta.$$

Thus,

$$t^2 = \frac{(b-a)^2\log(1/\delta)}{2n}.$$

### 7.2.4

We have shown that the uniform deviation between the population expectation and the empirical average is bounded as follows:

$$\sup_{f\in F}\left(\mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right) \le \mathbb{E}_n\left(\sup_{f\in F}\mathbb{E}f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}},$$

where:

- $\mathbb{E}f(z)$ is the population expectation of $f$,
- $\frac{1}{n}\sum_{i=1}^{n} f(z_i)$ is the empirical average over the sample $\{z_i\}_{i=1}^{n}$,
- $(b-a)$ is the range of the function values $f(z) \in [a,b]$,
- $\delta$ is the confidence parameter.

As such, we need to show that

$$\mathbb{E}_n\left(\sup_{f\in\mathcal{F}}\mathbb{E}f(z) - \tfrac{1}{n}\sum_{i=1}^{n} f(z_i)\right) \le \tfrac{2}{n}\mathbb{E}_n\mathrm{URad}(\mathcal{F}).$$

With the aforementioned results, we have

$$\sup_{f\in F}\mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \le \frac{2}{n}\,\mathrm{URad}\Big(F + 3(b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big).$$

**Proof.**
*Step 1: Bounding the Expected Supremum.*
We first show that the expected supremum can be bounded in terms of the Rademacher complexity:

$$\begin{aligned}
\mathbb{E}_n\left(\sup_f \mathbb{E}f(z) - \hat{\mathbb{E}}_n f\right) &\le \mathbb{E}_n\left(\mathbb{E}f^* - \hat{\mathbb{E}}_n f^* + \varepsilon\right) \\
&= \mathbb{E}_n\left(\mathbb{E}_n'\hat{\mathbb{E}}_n' f^* - \hat{\mathbb{E}}_n f^*\right) + \varepsilon \\
&= \mathbb{E}_n\mathbb{E}_n'\left(\hat{\mathbb{E}}_n' f^* - \hat{\mathbb{E}}_n f^*\right) + \varepsilon \\
&\le \mathbb{E}_n\mathbb{E}_n'\sup_f\left(\hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f\right) + \varepsilon.
\end{aligned}$$

Then, we have

$$\begin{aligned}\mathbb{E}_n\mathbb{E}_n' \sup_f \left(\hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f\right) &= \mathbb{E} \sup_f \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(f(z_i') - f(z_i)\right) \\ &\le \mathbb{E}_\varepsilon \mathbb{E}_n \mathbb{E}_n' \sup_{f,f'} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(f(z_i') - f'(z_i)\right) \\ &= \mathbb{E}_\varepsilon \mathbb{E}_n' \sup_f \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(z_i') + \mathbb{E}_\varepsilon \mathbb{E}_n \sup_{f'} \frac{1}{n}\sum_{i=1}^n (-\varepsilon_i) f'(z_i) \\ &= \frac{2}{n}\mathrm{URad}(\mathcal{F}).\end{aligned}$$

*Step 2: Combining the Results.*
With the aforementioned results, we obtain the final uniform convergence bound:

$$\begin{aligned}&\sup_{f\in F}\left(\mathbb{E}_z f(z) - \frac{1}{n}\sum_{i=1}^n f(z_i)\right) \\ &\le \frac{2}{n}\,\mathrm{URad}(F) + \frac{2}{n}(b-a)\sqrt{\frac{n\log(1/\delta)}{2}} + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \\ &= \frac{2}{n}\,\mathrm{URad}\Big(F + 3(b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big).\end{aligned}$$

### 7.2.5

Let $\ell \circ F_{|S}$ denote the function class defined as

$$(\ell \circ f)(x,y) = \ell\big(-yf(x)\big),$$

and define

$$\ell \circ F = \{\ell \circ f : f \in F\}.$$

Let $\ell : \mathbb{R}^n \to \mathbb{R}^n$ be a vector of univariate $L$-Lipschitz functions. Then,

$$\mathrm{URad}(\ell \circ V) \le L \cdot \mathrm{URad}(V).$$

**Proof.**
The idea of the proof is to 'de-symmetrize' and obtain a difference of coordinates to which we can apply the definition of the Lipschitz constant $L$. To start, we have

$$\begin{aligned}\mathrm{URad}(\ell \circ V) &= \mathbb{E}\sup_{u\in V}\sum_i \epsilon_i \ell_i(u_i) \\ &= \frac{1}{2}\,\mathbb{E}_{\epsilon_{2:n}} \sup_{u,w\in V}\Big(\ell_1(u_1) - \ell_1(w_1) + \sum_{i=2}^n \epsilon_i\big(\ell_i(u_i) + \ell_i(w_i)\big)\Big) \\ &\le \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{u,w\in V}\left(L|u_1 - w_1| + \sum_{i=2}^n \epsilon_i\left(l_i(u_i) + l_i(w_i)\right)\right).\end{aligned}$$

To get rid of the absolute value for any $\epsilon$ via a swapping argument between $u$ and $w$, we eventually obtain:

$$\begin{aligned}
&\sup_{u,w\in V}\left(L|u_1-w_1|+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right)\\
&=\max\left\{\sup_{u,w\in V}\left(L(u_1-w_1)+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right),\ \sup_{u,w\in V}\left(L(w_1-u_1)+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right)\right\}\\
&=\sup_{u,w\in V}\left(L(u_1-w_1)+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right).
\end{aligned}$$

As such,

$$\begin{aligned}
\mathrm{URad}(\ell\circ V)&\leq \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}}\sup_{u,w\in V}\left(L|u_1-w_1|+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right)\\
&=\frac{1}{2}\mathbb{E}_{\epsilon_{2:n}}\sup_{u,w\in V}\left(L(u_1-w_1)+\sum_{i=2}^n \epsilon_i(\ell_i(u_i)+\ell_i(w_i))\right)\\
&=\mathbb{E}_\epsilon\sup_{u\in V}\left[L\epsilon_1u_1+\sum_{i=2}^n\epsilon_i\ell_i(u_i)\right].
\end{aligned}$$

Repeating this procedure for the remaining coordinates gives the bound:

$$\mathrm{URad}(\ell\circ V)\leq \mathbb{E}_\epsilon\sup_u\left(L\sum_{i=1}^n\epsilon_iu_i\right)=L\cdot\mathrm{URad}(V).$$

Revisiting our overloaded composition notation:

$$(\ell\circ f)=\big((x,y)\mapsto\ell(-yf(x))\big),$$

and hence,

$$\ell\circ F=\{\ell\circ f: f\in F\}.$$