COMP 7070 Advanced Topics in AI and ML February 27, 2024

Lecture 7: Generalization Error

Instructor: Yifan Chen *Scribes: Hongduan Tian, Yi Ding* *Proof reader: Zhanke Zhou*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

7.1 Error Decomposition

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

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

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.

$$\mathcal{R}(\hat{f}) = \underbrace{\left[\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\right]}_{i.generalization} + \underbrace{\left[\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\right]}_{ii.optimization} + \underbrace{\left[\hat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f})\right]}_{iii.concentration/generalization} + \underbrace{\mathcal{R}(\bar{f})}_{iv.approximationerror} . \quad (7.1)$$

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

7.2 Generalization

Consider an infinite size of function space $\mathcal{F}$ o.w. we can use 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 ε-net.

7.2.1 Un-normalized Rademacher Complexity

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

$$\mathrm{URad}(\mathcal{V}) := \mathbb{E}_{\varepsilon}\sup_{a\in\mathcal{V}} < a, \varepsilon > . \quad (7.2)$$

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

7.2.2

Assume that $\mathcal{V} = \{(\ell(f(x_1), y_1), \ell(f(x_2), y_2), ..., \ell(f(x_n), y_n)) : f \in \mathcal{F}\}$, by applying Rademacher complexity to a dataset $\mathcal{S} = \{s_i = (x_i, y_i\}_{i=1}^n$, then the Rademacher complexity of $\ell \circ \mathcal{F}_{|\mathcal{S}}$ is:

$$\text{Urad}(\ell \circ \mathcal{F}_{|\mathcal{S}}) = \mathbb{E}_\varepsilon \text{sup}_{u \in \ell \circ \mathcal{F}_{|\mathcal{S}}} < \varepsilon, u >= \mathcal{E}_\varepsilon \text{sup}_{f \in \mathcal{F}} \left[\sum_{i=1}^n \varepsilon_i \ell(f(x_i), y_i) \right].$$

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

7.2.3

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

$$\begin{aligned} \mathbb{E}_z f(z) - \frac{1}{n} \sum_{i=1}^n f(z_i) &\leq \text{sup}_{f \in \mathcal{F}} \mathbb{E} f(z) - \frac{1}{n} \sum_{i=1}^n f(z_i) \\ &\leq \mathbb{E}_{z_i} \left(\text{sup}_{f \in \mathcal{F}} \mathbb{E} f(z) - \frac{1}{n} \sum_{i=1}^n f(z_i) \right) + (b - a) \cdot \sqrt{\frac{\log 1/\delta}{2n}}. \end{aligned}$$

Remark. Based on the above summary, we can further have $\left| \text{sup}_{f \in \mathcal{F}} \mathbb{E} f(z) - \frac{1}{n} \sum_{i=1}^n f(z_i) - \mathbb{E}_{z_i} \right| \leq 2(b - a) \sqrt{\frac{\log 1/\delta}{2n}}$.

Lemma 7.1. *Given two functions f, g, we have* $\text{sup}_a f(a) + g(a) \leq \text{sup}_{a^*} (\text{sup}_a f(a) + g(a^*))$.

Proof of Lemma 7.1

Proof. Firstly, we know that $\forall \varepsilon, \exists a^*$ that satisfies

$$\text{sup}_a f(a) + g(a) \leq f(a^*) + g(a^*) + \varepsilon.$$

Then, we know

$$\text{sup}_a f(a) + g(a^*) \geq f(a^*) + g(a^*) \geq \text{sup}_a f(a) + g(a) - \varepsilon.$$

Thus,

$$RHS = \text{sup}_{a^*} (\text{sup}_a f(a) + g(a^*)) \geq \text{sup}_a f(a) + g(a).$$

□

Lemma 7.2. *Given two functions f, g, we have* $-\text{sup}_a (f(a) + g(a)) \leq \text{sup}_{a^*} (-\text{sup}_a f(a) - g(a^*))$.

Proof of Remark

Proof. Firstly, we know that $\text{sup}_{f \in \mathcal{F}} \mathbb{E} f(z) - \frac{1}{n} \sum_{i=1}^n f(z_i) \sim \mathsf{subG}$ and it is a function of $z \sim z_n$ with the bounded difference. Then, we have

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

where $z_i^{\setminus j} = \{z_1, z_2, ..., z_n\}$. For convenience, we reload the notations as following

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

Then, Eq.(7.3) can be further formulated as:

$$\begin{aligned}
&\left|\sup_f\left(\mathbb{E}_z f(z)-\frac{1}{n}\sum_{i=1}^n f(z_i)\right)-\sup_{f'}\left(\mathbb{E}_z f'(z)-\frac{1}{n}\sum_{i=1}^n f'(z_i^{\setminus j})\right)\right| \\
&=\left|\sup_f\left(\mathbb{E}f-\hat{\mathbb{E}}f\right)-\sup_f'\left(\mathbb{E}f'-\hat{\mathbb{E}}_n f'+\frac{1}{n}f'(z_j)-\frac{1}{n}f'(z_j')\right)\right| \\
&\le \sup_{f''}\left|\sup_f(\mathbb{E}_f-\hat{\mathbb{E}}_n f)-\sup_{f'}(\mathbb{E}f'-\hat{\mathbb{E}}_n f')-\frac{1}{n}\left(f''(z_i)-f''(z_i')\right)\right|
\end{aligned} \tag{7.4}$$

With Lemma 7.2, we have

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

Similarly, we have

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

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

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

Thus,

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

□

7.2.4

Have shown that $\sup_{f\in\mathcal{F}}\mathbb{E}f(z)-\frac{1}{n}\sum_{i=1}^n f(z_i)\le\mathbb{E}_n\left(\sup_{f\in\mathcal{F}}\mathbb{E}f(z)-\frac{1}{n}\sum_{i=1}^n f(z_i)\right)+(b-a)\sqrt{\frac{\log 1/\delta}{2n}}$, we need to further show $\mathbb{E}_n\left(\sup_{f\in\mathcal{F}}\mathbb{E}f(z)-\frac{1}{n}\sum_{i=1}^n f(z_i)\right)\le\frac{2}{n}\mathbb{E}_n\mathrm{URad}(\mathcal{F})$.

With the aforementioned results, we have

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

Proof. Firstly,

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

Then,

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

□

7.2.5

$\ell\circ\mathcal{F}_{|\mathcal{S}}$: Let $\ell:\mathbb{R}^n\to\mathbb{R}^n$ be a vector of univariate L-lipschitz functions. Then $\mathrm{URad}(\ell\circ V)\leq L\cdot\mathrm{URad}(V)$.

Proof. The idea of the proof is to 'de-symmetrize' and get a difference of coordinates to which we can apply the definition of L. To start,

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

To get rid of the absolute value, for any ϵ, by considering swapping u and w,

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

As such,

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

Repeating this procedure for the other coordinates gives the bound.

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

□

Revisiting our overloaded composition notation:

$$(\ell \circ f) = ((x, y) \mapsto \ell(-yf(x))),$$
$$\ell \circ \mathcal{F} = \{\ell \circ f : f \in \mathcal{F}\}.$$