6 Rademacher Complexity: Definition

So far, our quest to achieve generalization involves three key steps: 1-concentration, 2- union bound, and 3- uniform conv $\implies$ generalization. That is, we used Hoeffding's lemma to obtain a concentration result for the empirical risk. We then establish that an empirical process is small, by either handling the supremum through a union bound over either a finite function class, or using an ε-net argument to obtain a generalization bound. Lastly, using that unif. conv $\implies$ generalization, we get

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\Bigg(\underbrace{\sup_{f\in\mathcal{F}} \; |\hat{R}(f) - R(f)| \geq \frac{\epsilon}{2}}_{\text{empirical process}}\Bigg) = (*) \tag{6.1}$$

$$\leq \text{ func. of } \Bigg(\epsilon, \text{ complexity of } \mathcal{F}, n\Bigg).$$

In the following, all of the above steps will be modified– steps 1 and 3 change slightly, but 2 entirely. In the concentration step, we will obtain a concentration result directly for the empirical process (not for empirical risk), showing that it is close to its expectation. Then we will use a technique called "symmetrization" to show that the expectation of the empirical process depends on the complexity of the function class. Lastly, by using a slightly modified version of "uniform conv. $\implies$ generalization" we will obtain a generalization bound.

Rademacher complexity of the function class $\mathcal{F}$ over n samples will be denoted by $\mathfrak{R}_n(\mathcal{G})$. We will be formally defining the Rademacher complexity **later** in this section, but in the sequel it should be understood as a measure of complexity of the function class $\mathcal{G}$ over n data points.

6.1 Generalization based on Rademacher complexity

Theorem 23 (Generalization based on Rademacher complexity)**.** *Define*

$$\mathcal{G} = \{(y, x) \to \ell((y, x), f) \text{ where } f \in \mathcal{F}\},$$

and assume ℓ is bounded, $\ell \in [0, B]$, and $(x_i, y_i) \overset{iid}{\sim} p$. Then with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f) \leq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{6.2}$$

Before proving the above theorem, we make a few remarks.

Remarks

- As stated before, Rademacher complexity measures the complexity of the function class $\mathcal{G}$ over n data points. It should converge to zero as n gets large, and this determines the generalization error rate.
- It is important to note that in the bound (6.2), $\mathfrak{R}_n(\mathcal{G})$ is the Rademacher complexity of the function class $\mathcal{G}$, not $\mathcal{F}$. We will connect this to $\mathcal{F}$ later.
- Although what we care about is bounding the generalization error, the above bound is obtained for the empirical process, and the technique used here has applications beyond generalization.

Proof. Our proof strategy is as follows.

We will conclude our proof with a modified version of the "uniform conv. $\implies$ generalization". For this, we start by splitting the inequality (6.1) into two components:

$$* \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \ \hat{R}(f) - R(f) \geq \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}} \ R(f) - \hat{R}(f) \geq \frac{\epsilon}{2} \right). \tag{6.3}$$

In the remainder of this proof, we will focus on bounding the first term on the right hand side above. But an equivalent bound can be shown for the second term.

The proof relies on three key steps as before: 1-concentration, 2-symmetrization, and 3- uniform conv. $\implies$ generalization.

1. **Concentration**: Previously, we relied on Hoeffding's inequality to obtain a concentration bound for the empirical risk. In the sequel, we will use a stronger theorem in order to obtain a concentration result directly for the empirical process.

   Lemma 24 (McDiarmid's inequality)**.** *Let $g : \mathcal{Z} \times ... \times \mathcal{Z} \to \mathbb{R}$ be a function satisfying the bounded difference property*

   $$|g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z_j', \ldots, z_n)| \leq c_j$$

   Then for independent random variables $z_1, z_2 \ldots, z_n$, we have

   $$\mathbb{P}\Big(g(z_1, \ldots, z_n) - \mathbb{E}[g(z_1, \ldots, z_n)] \geq \epsilon \Big) \leq \exp\left\{ \frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right\}.$$

   This lemma is stronger than the previously used concentration arguments. Indeed, Hoeffding's inequality can be derived by using the above lemma.

   Example: [Hoeffding's inequality] Suppose $z_1, \ldots, z_n$ are independent random variables that are bounded almost surely $a_i \leq z_i \leq b_i$. We define g as their average and verify the bounded difference property

   $$\begin{aligned} g(z_1, ..., z_n) = S_n &= \frac{1}{n} \sum_{i=1}^n z_i \\ |g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z_j', \ldots, z_n)| &\leq \frac{1}{n} |z_j - z_j'| \\ &\leq \frac{b_j - a_j}{n}. \end{aligned}$$

   By the McDiarmid's inequality, we obtain

   $$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n z_i - \mathbb{E}[z_1] \geq \epsilon \right) \leq \exp\left\{ \frac{-2\epsilon^2 n}{\sum_j (b_j - a_j)^2} \right\}.$$

   We continue the proof by recalling our goal: We need to bound the empirical process in (6.3).

For this, we let the g function from McDiarmid's inequality be the function of interest.

$$\begin{aligned} g(z_1, \ldots, z_n) &= \sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(\underbrace{(x_i, y_i)}_{z_i}, f) - \mathbb{E}[\ell(\underbrace{(x, y)}_{z}, f)] \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] \end{aligned}$$

Notice that, in order to ease the notation, we denoted the data pairs (x_i, y_i) as z_i. We first verify the bounded difference property.

$$\begin{aligned} &|g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z'_j, \ldots, z_n)| \\ &= \left| \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] \right] - \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] - \frac{1}{n} \Big\{ \ell(z_j, f) - \ell(z'_j, f) \Big\} \right] \right| \\ &\underbrace{\leq}_{\text{(i)}} \sup_{f \in \mathcal{F}} \frac{1}{n} \big| \ell(z_j, f) - \ell(z'_j, f) \big| \\ &\leq \frac{B}{n} \end{aligned}$$

where the inequality (i) follows from the following simple fact. For function F, G, we have

Fact 25.

$$\Big| \sup_x F(x) - \sup_x G(x) \Big| \leq \sup_x \Big| F(x) - G(x) \Big|.$$

Hence, by the McDiarmid's inequality, we obtain

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \geq t + \overbrace{\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \Big]}^{\text{Need to show is small}} \right) \underbrace{\leq}_{\text{McDiarmid's}} \exp\left\{ \frac{-2nt^2}{B^2} \right\}.$$

It is worth highlighting the following again. Previously, we have focused on the concentration over $n^{-1} \sum_i \ell(z_i, f) - E[\ell(z, f)]$, followed by a union bound, but now we are looking at the concentration of the supremum directly $\sup_{f \in \mathcal{F}} \; n^{-1} \sum_i \ell(z_i, f) - E[\ell(z, f)]$. The above bound is looking good for our goal except that we need to control the additional term that is the expected value of the empirical process. This will be done by the symmetrization argument.

6.2 Symmetrization

We start with a simple argument. If X, X' are iid r.v.'s then $X \overset{d}{=} X'$. If g is a function then $g(X) \overset{d}{=} g(X')$. Further,

$$\begin{aligned} g(X) - g(X') &\overset{d}{=} g(X') - g(X) \\ &\overset{d}{=} -1 \cdot [g(X) - g(X')] \\ &\overset{d}{=} \sigma \cdot (g(X) - g(X')), \end{aligned}$$

where σ is a Rademacher random variable, i.e. $\mathbb{P}(\sigma = +1) = \mathbb{P}(\sigma = -1) = 1/2$, which is independent of X, X'. This argument is very useful (as we will see soon), and termed as symmetrization.

2. **Symmetrization**: Denote our dataset as: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} = \{z_1, \ldots, z_n\}$. Introduce a random copy of the dataset $\mathcal{D}' = \{z'_1, \ldots, z'_n\}$, where z_i and z'_i are iid. This new dataset is called the ghost dataset. Now that we have two datasets $\mathcal{D}, \mathcal{D}'$, there are also two empirical risks $R(f; \mathcal{D})$ and $R(f; \mathcal{D}')$ where we denote their dependence on the corresponding dataset. That is,

$$\hat{R}(f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) \quad \text{and} \quad \hat{R}(f; \mathcal{D}') = \frac{1}{n} \sum_{i=1}^{n} \ell(z'_i, f).$$

The population risk will be identical for these datasets,

$$R(f) = E[\ell(z, f)] = E[\hat{R}(f, \mathcal{D})] = E[\hat{R}(f, \mathcal{D}')].$$

We write,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f) - R(f)\right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f; \mathcal{D}) - \mathbb{E}[\hat{R}(f; \mathcal{D}')]\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \Big\{\hat{R}(f; \mathcal{D}) - \underbrace{\mathbb{E}[\hat{R}(f; \mathcal{D}')|D]}_{D \overset{\text{indep}}{\sim} D'}\Big\}\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \Big\{\mathbb{E}\Big[\hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')|D\Big]\Big\}\right] \\
&\underbrace{\leq}_{\text{(i)}} \mathbb{E}\left[\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \; \Big\{\hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')\Big\}|D\Big]\right] \\
&\underbrace{=}_{\text{(ii)}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \Big(\ell(z_i, f) - \ell(z'_i, f)\Big)\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(\ell(z_i, f) - \ell(z'_i, f)\Big)\right] \\
&\underbrace{\leq}_{\text{(iii)}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) + \sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} -\sigma_i \ell(z'_i, f)\right] \\
&\underbrace{=}_{\sigma_i \overset{d}{=} -\sigma_i} 2 \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f)\right]
\end{aligned}$$

where (i) follows from $\mathbb{E}[\sup] \geq \sup \mathbb{E}$, (ii) follows from the law of iterated expectation, and (iii) follows from the following fact.

Fact 26. $\sup_x \{F(x) + G(x)\} \leq \sup_x F(x) + \sup_x G(x)$.

The bound we obtained through the above steps is simply,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(z_i, f)\right], \tag{6.4}$$

and the final bound doesn't include the ghost dataset at all!

Next, we define the Rademacher complexity.

Definition 27 (Rademacher complexity)**.** *For a function class $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}\}$, the Rademacher complexity is defined as,*

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)\right],$$

where $z_i \overset{iid}{\sim} p$ are data points, and $\sigma_i \overset{iid}{\sim}$ Rademacher r.v.'s independent from the dataset. Furthermore, the empirical Rademacher complexity is defined as,

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i) | z_{1:n}\right].$$

Therefore, defining the function class $\mathcal{G}$ as

$$\mathcal{G} = \{g : z \to \ell(z, f) \text{ such that } f \in \mathcal{F}\},$$

the bound in (6.4) can be written as

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2\mathfrak{R}_n(\mathcal{G}).$$

3. **Uniform convergence $\implies$ generalization** (yet again): We now have the necessary building blocks to construct our goal of generalization. We write out generalization bound as

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2}\right).$$

We will obtain (already obtained) a bound on the first term on the right hand side. Similar argument yields the same bound for the second term.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq t + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\Big]\right) \leq \exp\left\{\frac{-2nt^2}{B^2}\right\}$$

Using that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2 \cdot \mathfrak{R}_n(\mathcal{G})$$
$$\text{where } \mathcal{G} = \{z \to \ell(z, f) \text{ where } f \in \mathcal{F}\},$$

we can write,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \underbrace{t + 2\mathfrak{R}_n(\mathcal{G})}_{\epsilon/2} \right) \leq \underbrace{\exp\left\{\frac{-2nt^2}{B^2}\right\}}_{\delta}$$

Then with probability at least $1-\delta \coloneqq 1 - 2\exp\left\{\frac{-2nt^2}{B^2}\right\}$, we have

$$\hat{R}(f) - R(f) \leq \epsilon/2 = t + 2\mathfrak{R}_n(\mathcal{G})$$
$$\text{for} \quad t = B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Similar argument holds for $\mathbb{P}\Big(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2}\Big)$. Therefore, we obtain

$$\mathbb{P}\Big(R(\hat{f}) - R(f^*) \geq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}\Big) \leq 1-\delta,$$

which concludes the proof.

□

Outline of the above proof is as follows.

1. Concentration of the empirical process

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \leq t + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\Big] \right) \underbrace{\leq}_{\text{McDiarmid's}} \exp\Big\{\frac{-2nt^2}{B^2}\Big\}.$$

2. Symmetrization: For $\mathcal{G} = \{z \to \ell(z, f) \text{ where } f \in \mathcal{F}\}$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \right] \leq 2 \cdot \mathbb{E}\left[\sup_{f\in\mathcal{F}} n^{-1} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) \right] = 2 \cdot \mathfrak{R}_n(\mathcal{G}).$$

3. Uniform convergence implies generalization

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2} \right).$$

where we set $\epsilon/2 = t + 2\mathfrak{R}_n(\mathcal{G})$ and $\delta = \exp\left\{\frac{-2nt^2}{B^2}\right\}$ and solve for these quantities.