CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: Will Song

Lecture # 5
Jan 27th, 2021

5.1 Review and overview

Recall that our goal is to bound the *excess risk* $L(\hat{h}) - L(h^*)$, where L is the expected loss (or population loss), $\hat{h}$ is our estimated hypothesis and h^* is the hypothesis in the hypothesis class $\mathcal{H}$ which minimizes the expected loss. In previous lectures, we showed that to do so it suffices to upper bound $\sup_{h\in\mathcal{H}}(L(h) - \widehat{L}(h))$. (Note: we often call $L(\hat{h}) - \widehat{L}(\hat{h})$ the *generalization gap* or *generalization error*.)

In the previous lecture, we derived bounds for the generalization gap in two cases:

1. If the hypothesis class $\mathcal{H}$ is finite,

$$L(\hat{h}) - \hat{L}(\hat{h}) \leq \tilde{O}\left(\sqrt{\frac{\log|\mathcal{H}|}{n}}\right). \tag{5.1}$$

2. If the hypothesis class $\mathcal{H}$ is p-dimensional,

$$L(\hat{h}) - \hat{L}(\hat{h}) \leq \tilde{O}\left(\sqrt{\frac{p}{n}}\right). \tag{5.2}$$

Both of these bounds have a $\frac{1}{\sqrt{n}}$-dependency on n, which is known as the "slow rate". The terms in the numerator ($\log|\mathcal{H}|$ and p resp.) can be thought of as complexity measure of $\mathcal{H}$.

In this lecture we will introduce a more general complexity measure (*Rademacher complexity*) that is mathematically cleaner and gives tighter bounds, and in the next lecture we will instantiate this complexity measure for special hypothesis classes.

5.2 Motivation for a new complexity measure

The bound (5.2) is not precise enough: it depends solely on p and is not always optimal. For example, this would be a poor bound if the hypothesis class $\mathcal{H}$ has very high dimension but small norm. One specific example is for the following two hypothesis classes:

$$\{\theta : \|\theta\|_1 \leq B\} \qquad \text{vs.} \qquad \{\theta : \|\theta\|_2 \leq B\},$$

(5.2) would give both hypothesis classes the same bound of $\tilde{O}\left(\sqrt{\frac{p}{n}}\right)$. Intuitively, we should take into account the norms for a better bound.

With the complexity measure to be introduced, we will prove a bound of the form

$$L(\hat{h}) - \hat{L}(\hat{h}) \leq \tilde{O}\left(\sqrt{\frac{\text{Complexity}(\Theta)}{n}}\right). \tag{5.3}$$

This complexity measure will depend on the distribution p over $\mathcal{X} \times \mathcal{Y}$ (the input and output spaces), and hence takes into account how easy it is to learn p. If p is easy to learn, then this complexity measure will be small even if the hypothesis space is big.

One of the practical implications of having such a complexity measure is that we can restrict the hypothesis space by regularizing the complexity measure (assuming it is something we can evaluate and train with). If we successfully find a low complexity model, then this generalization bound guarantees that we have not overfit.

5.3 Rademacher complexity

In uniform convergence, we sought a high probability bound for $\sup_{h \in H}(L(h) - \hat{L}(h))$. Here we have a weaker goal: we try to obtain an upper bound for its expectation instead, i.e.

$$\mathbb{E}_{z \sim X \times Y}\left[\sup_{h \in H}(L(h) - \hat{L}(h))\right] \leq \text{ upper bound.} \tag{5.4}$$

The expectation is over the randomness in the training data ($\mathcal{X} \times \mathcal{Y}$). (Note: We cannot just swap the order of $\mathbb{E}$ and sup!)

To do so, we first define *Rademacher complexity*.

Definition 5.1 (Rademacher complexity)**.** Let $\mathcal{F}$ be a family of functions mapping $Z \mapsto \mathbb{R}$, and let P be a distribution over Z. The *(average) Rademacher complexity* of $\mathcal{F}$ is defined as

$$R_n(F) \triangleq \mathbb{E}_{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P}\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n \overset{\text{iid}}{\sim} \{\pm 1\}}\left[\sup_{f \in F} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right]\right], \tag{5.5}$$

where $\sigma_1, \ldots, \sigma_n$ are independent *Rademacher random variables*, i.e. each taking on the value of 1 or -1 with probability $1/2$.

Remark 5.2. For our applications we will take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. However, Definition 5.1 holds for abstract input spaces $\mathcal{Z}$ as well.

Remark 5.3. Note that $R_n(\mathcal{F})$ is also dependent on the measure P of the space, so technically it should be $R_{n,P}(\mathcal{F})$, but for brevity, we refer to it as $R_n(\mathcal{F})$.

An interpretation is $R_n(\mathcal{F})$ is the maximal possible correlation between outputs of some $f \in \mathcal{F}$ (on points $f(z_1), \ldots, f(z_n)$) and random Rademacher variables $(\sigma_1, \ldots, \sigma_n)$. Essentially, functions with more random sign outputs will better match random patterns of Rademacher variables and have higher complexity (greater ability to mimic or express randomness).

The following theorem is the main theorem involving Rademacher complexity:

Theorem 5.4.

$$\mathbb{E}_{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P}\left[\sup_{f \in F}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z \sim P} f(x)\right]\right] \leq 2R_n(\mathcal{F}). \tag{5.6}$$

Remark 5.5. We can think of $\frac{1}{n}\sum_{i=1}^{n} f(z_i)$ as an empirical average and $\mathbb{E}_{z \sim P} f(x)$ as a population average.

Why is Theorem 5.4 useful to us? We can set $\mathcal{F}$ to be the family of loss functions, i.e.

$$\mathcal{F} = \{z = (x, y) \in \mathcal{Z} \mapsto \ell((x, y), h) \in \mathbb{R} : h \in \mathcal{H}\}. \tag{5.7}$$

This is the family of losses induced from the hypothesis functions in $\mathcal{H}$. Then by Theorem 5.4,

$$\mathbb{E}_{z_i}\left[\sup_{f\in F}\left(\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}(f)\right)\right] = \mathbb{E}_{(x^{(i)},y^{(i)})}\left[\sup_{h\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell((x^{(i)}, y^{(i)}, h)) - L(h)\right]\right] \tag{5.8}$$

$$= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\hat{L}(h) - L(h)\right)\right] \tag{5.9}$$

$$\leq 2R_n(\mathcal{F}), \tag{5.10}$$

where $z_i = (x^{(i)}, y^{(i)})$. Thus, $2R_n(\mathcal{F})$ is an upper bound for the generalization error. In this context, $R_n(\mathcal{F})$ can be interpreted as how well the loss sequence $\ell((x^{(1)}, y^{(1)}), h), \ldots \ell((x^{(n)}, y^{(n)}), h)$ correlates with $\sigma_1, \ldots, \sigma_n$.

Example 5.6. Consider the binary classification setting where $y \in \{\pm 1\}$. Let ℓ_{0-1} denote the zero-one loss function. Note that

$$\ell_{0-1}((x, y), h) = \mathbf{1}\{h(x) \neq y\} = \frac{1 - yh(x)}{2}. \tag{5.11}$$

Hence,

$$R_n(\mathcal{F}) = \mathbb{E}_{(x^{(i)},y^{(i)}),\sigma_i}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\ell_{0-1}((x^{(i)}, y^{(i)}), h)\sigma_i\right] \quad \text{(by definition)} \tag{5.12}$$

$$= \mathbb{E}_{(x^{(i)},y^{(i)}),\sigma_i}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{-h(x^{(i)})y^{(i)} + 1}{2}\right)\sigma_i\right] \quad \text{(by (5.11))} \tag{5.13}$$

$$= \frac{1}{2}\mathbb{E}_{(x^{(i)},y^{(i)}),\sigma_i}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i + \sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n} -h(x^{(i)})y^{(i)}\sigma_i\right] \quad \text{(sup only over } \mathcal{H}) \tag{5.14}$$

$$= \frac{1}{2}\mathbb{E}_{(x^{(i)},y^{(i)}),\sigma_i}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n} -h(x^{(i)})y^{(i)}\sigma_i\right] \quad (\mathbb{E}[\sigma_i] = 0) \tag{5.15}$$

$$= \frac{1}{2}\mathbb{E}_{(x^{(i)},y^{(i)}),\sigma_i}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})\sigma_i\right] \quad (-y_i\sigma_i \overset{d}{=} \sigma_i) \tag{5.16}$$

$$= \frac{1}{2}R_n(\mathcal{H}). \quad \text{(by definition)} \tag{5.17}$$

In this setting, $R_n(\mathcal{F})$ and $R_n(\mathcal{H})$ are the same (except for the factor of 2). $R_n(\mathcal{H})$ has a slightly more intuitive interpretation: it represents how well $h \in \mathcal{H}$ can fit random patterns.

Warning! $R_n(\mathcal{F})$ is not always the same as $R_n(\mathcal{H})$ in other problems.

Remark 5.7. Rademacher complexity is invariant to translation. One example of this in play is in how the $+1$ in the $\left(\frac{-h(x^{(i)})y^{(i)}+1}{2}\right)$ term essentially vanishes in the computation.

Let us now prove Theorem 5.4.

Proof of Theorem 5.4. We use a technique called *symmetrization*, which is a very important technique in probability theory. We first fix $z_1, \ldots, z_n$and draw $z'_1, \ldots z'_n \overset{\text{iid}}{\sim} P$. Then we can rewrite the term in the expectation on the LHS of (5.6):

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f]\right) = \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}_{z'_1,\ldots,z'_n}\left[\frac{1}{n}\sum_{i=1}^n f(z'_i)\right]\right) \tag{5.18}$$

$$= \sup_{f\in\mathcal{F}}\left(\mathbb{E}_{z'_1,\ldots,z'_n}\left[\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(z'_i)\right]\right) \tag{5.19}$$

$$\leq \mathbb{E}_{z'_1,\ldots,z'_n}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(z'_i)\right)\right]. \tag{5.20}$$

The last inequality is because in general,

$$\sup_u\left(\mathbb{E}_v[g(u,v)]\right) \leq \sup_u\left(\mathbb{E}_v\left[\sup_{u'} g(u',v)\right]\right) = \mathbb{E}_v\left[\sup_u(g(u,v))\right] \tag{5.21}$$

since the outer sup becomes vacuous.

Now, if we take the expectation over $z_1, \ldots, z_n$ for both sides of (5.20),

$$\mathbb{E}_{z_1,\ldots,z_n}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f]\right)\right] \leq \mathbb{E}_{z_i}\left[\mathbb{E}_{z'_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \left(f(z_i) - f(z'_i)\right)\right)\right]\right] \tag{5.22}$$

$$= \mathbb{E}_{z_i,z'_i}\left[\mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(z_i) - f(z'_i)\right)\right)\right]\right] \tag{5.23}$$

$$\leq \mathbb{E}_{z_i,z'_i,\sigma_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right) + \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n -\sigma_i f(z'_i)\right)\right] \tag{5.24}$$

$$= 2R_n(\mathcal{F}), \tag{5.25}$$

where (5.23) is because $\sigma_i(f(z_i)-f(z'_i)) \overset{d}{=} f(z_i)-f(z'_i)$ since $f(z_i)-f(z'_i)$ has a symmetric distribution. The last equality holds since $-\sigma_i \overset{d}{=} \sigma_i$ and z_i, z'_i are drawn iid from the same distribution. □

Here is an intuitive understanding of what Theorem 5.4 achieves. Consider the quantities on the LHS and RHS of (5.6):

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f]\right) \qquad \text{v.s.} \qquad \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right).$$

First, we removed $\mathbb{E}[f]$, which is hard to control quantitatively since it is deterministic. Second, we added more randomness in the form of Rademacher variables. This will allow us to shift our focus from the randomness in the z_i's to the randomness in the σ_i's. In the future, our bounds on the Rademacher complexity will typically only depend on the randomness from the σ_i's.

5.3.1 Dependence of Rademacher complexity on P

For intuition on how Rademacher complexity depends on the distribution P, consider the extreme example where P is a point mass, i.e. $z = z_0$ almost surely. Assume that $-1 \leq f(z_0) \leq 1$ for all $f \in \mathcal{F}$. Then

$$\mathbb{E}_{z_1,\ldots,z_n \sim P}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n} f(z_0)\sum_{i=1}^{n} \sigma_i\right] \tag{5.26}$$

$$\leq \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i\right|\right] \qquad \text{(since } f(z_0) \in [-1,1]\text{)} \tag{5.27}$$

$$\leq \mathbb{E}_{\sigma_i}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i\right)^2\right]^{\frac{1}{2}} \qquad \text{(Jensen's Inequality)} \tag{5.28}$$

$$= \frac{1}{n}\left(\mathbb{E}_{\sigma_i,\sigma_j}\left[\sum_{i,j=1}^{n} \sigma_i\sigma_j\right]\right)^{\frac{1}{2}} \tag{5.29}$$

$$= \frac{1}{n}\left(\mathbb{E}_{\sigma_i}\left[\sum_{i=1}^{n} \sigma_i^2\right]\right)^{\frac{1}{2}} \tag{5.30}$$

$$= \frac{1}{n}\cdot\sqrt{n} = \frac{1}{\sqrt{n}}. \tag{5.31}$$

This bound does not depend on $\mathcal{F}$ (except that it is bounded). This example illustrates that sometimes we only need to depend on the distribution of Rademacher random variables.

5.4 Emperical Rademacher complexity

In the previous section, we bounded the expectation of $\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z\sim P} f(x)\right]$. This expectation is taken over the training examples $z_1, \ldots, z_n$. In many instances we only have one training set, and do not have access to many training sets. Thus, the bound on the expectation does not give a guarantee for the one training set that we have. In this section, we seek to bound the quantity itself with high probability.

Definition 5.8 (Empirical Rademacher complexity)**.** Given a dataset $S = \{z_1, \ldots, z_n\}$, the *emprical Rademacher complexity* is defined as

$$R_S(\mathcal{F}) \triangleq \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right]. \tag{5.32}$$

$R_S(\mathcal{F})$ is a function of both the function class $\mathcal{F}$ and the dataset S.

Note that, as the name suggests, the expectation of the empirical Rademacher complexity is the Rademacher complexity:

$$R_n(\mathcal{F}) = \mathop{\mathbb{E}}_{\substack{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P \\ S=\{z_1,\ldots,z_n\}}} \left[R_S(\mathcal{F})\right]. \tag{5.33}$$

Here is the theorem involving empirical Rademacher complexity:

Theorem 5.9. *Suppose for all $f \in \mathcal{F}$, $0 \le f(x) \le 1$. Then, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f]\right] \le 2R_S(F) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{5.34}$$

Proof. For conciseness, define

$$g(z_1,\dots,z_n) \triangleq \sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f]\right]. \tag{5.35}$$

We prove the theorem in 4 steps.

Step 1: We bound g using McDiarmid's Inequality. To use McDiarmid's inequality, we check that the bounded difference condition holds:

$$g(z_1,\dots,z_n) - g(z_1,\dots,z_i',\dots,z_n) \le \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{j=1}^{n} f(z_j)\right] - \sup_{f\in\mathcal{F}}\left[\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} f(z_j)\right) + \frac{f(z_i')}{n}\right] \tag{5.36}$$

$$\le \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\left(f(z_i) - f(z_i')\right)\right] \tag{5.37}$$

$$\le \frac{1}{n}. \tag{5.38}$$

(5.37) holds because in general, $\sup_f A(f) - \sup_f B(f) \le \sup_f [A(f) - B(f)]$, and (5.38) holds since f is bounded by $[0,1]$. We can thus apply McDiarmid's Inequality with parameters $c_1 = \cdots = c_n = 1/n$:

$$\mathbb{P}\left[g(z_1,\dots,z_n) \ge \mathbb{E}_{z_1,\dots,z_n \overset{\text{iid}}{\sim} P}[g] + \epsilon\right] \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp(-2n\epsilon^2). \tag{5.39}$$

Step 2: We apply Theorem 5.4 to get

$$\mathbb{E}_{z_1,\dots,z_n \overset{\text{iid}}{\sim} P}[g] \le 2R_n(\mathcal{F}). \tag{5.40}$$

Step 3: Define

$$\tilde{g}(z_1,\dots,z_n) = R_S(\mathcal{F}) \triangleq \mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]. \tag{5.41}$$

Using a similar argument like that in Step 1, we show that $\tilde{g}$ satisfies the bounded difference condition:

$$\tilde{g}(z_1,\dots,z_n) - \tilde{g}(z_1,\dots,z_i',\dots,z_n)$$

$$\le \mathbb{E}_{\sigma_i}\left[\sup_{f\in F}\left[\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j)\right] - \sup_{f\in F}\left[\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n}\sigma_j f(z_j)\right) + \frac{1}{n}\sigma_i f(z_i')\right]\right] \tag{5.42}$$

$$\le \mathbb{E}_{\sigma_i}\left[\sup_{f\in F}\left(\frac{1}{n}\sigma_i(f(z_i) - f(z_i'))\right)\right] \tag{5.43}$$

$$\le \frac{1}{n}, \tag{5.44}$$

since the term inside the sup is always upper bounded by 1. We can thus apply McDiarmid's Inequality with parameters $c_1 = \cdots = c_n = 1/n$:

$$\mathbb{P}\left[\tilde{g} - \mathbb{E}[\tilde{g}] \geq \epsilon\right] \leq \exp(-2n\epsilon^2), \quad \text{and} \quad \mathbb{P}\left[\tilde{g} - \mathbb{E}[\tilde{g}] \leq -\epsilon\right] \leq \exp(-2n\epsilon^2). \tag{5.45}$$

Step 4: We set δ such that $\exp(-2n\epsilon^2) = \delta/2$. (This implies that $\epsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$.) Then, with probability $\geq 1 - \delta$,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f]\right] = g &\leq \mathbb{E}[g] + \epsilon && \text{(Step 1)} && (5.46)\\
&\leq 2R_n(\mathcal{F}) + \epsilon && \text{(Step 2)} && (5.47)\\
&\leq 2(R_S(\mathcal{F}) + \epsilon) + \epsilon && \text{(Step 3)} && (5.48)\\
&= 2R_S(\mathcal{F}) + 3\epsilon, && && (5.49)
\end{aligned}$$

as required. □

Setting $\mathcal{F}$ to be a family of loss functions bounded by $[0, 1]$ in Theorem 5.9 gives the following corollary:

Corollary 5.10. Let $\mathcal{F}$ to be a family of loss functions $\mathcal{F} = \{(x, y) \mapsto \ell((x, y), h) : h \in \mathcal{H}\}$ with $\ell((x, y), h) \in [0, 1]$ for all ℓ, (x, y) and h. Then, with probability $1 - \delta$, the generalization gap is

$$L(h) - \hat{L}(h) \leq 2R_S(F) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{for all } h \in \mathcal{H}.$$