Chapter 4

Generalization Bounds via Uniform Convergence

In Chapter 2, we pointed out some limitations of asymptotic analysis. In this chapter, we will turn our focus to *non-asymptotic analysis*, where we provide convergence guarantees without having the number of observations n go off to infinity. A key tool for proving such guarantees is *uniform convergence*, where we have bounds of the following form:

$$\Pr\left[\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \le \epsilon\right] \ge 1-\delta. \tag{4.1}$$

In other words, the probability that the difference between our empirical loss and population loss is larger than ϵ is at most δ. We give motivation for uniform convergence and show how it can give us non-asymptotic guarantees on excess risk.

4.1 Basic concepts

A central goal of learning theory is to bound the *excess risk* $L(\hat{\theta}) - L(\theta^*)$. This is important as we don't want the expected risk of our ERM to be much larger than the expected risk of the best possible model. As we will see in the remainder of this section, uniform convergence is a technique that helps us achieve such bounds.

Uniform convergence is a property of a parameter set Θ, which gives us bounds of the form

$$\Pr\left[|\hat{L}(\theta) - L(\theta)| \ge \varepsilon\right] \le \delta;, \forall\theta\in\Theta. \tag{4.2}$$

In other words, uniform convergence tells us that for any choice of θ, our empirical risk is always close to our population risk with high probability. Let's look at a motivating example for why this type of bound is useful.

4.1.1 Motivation: Uniform convergence implies generalization

Consider the standard supervised learning setup where we have some i.i.d. $\{(x^{(i)}, y^{(i)})\}$. Furthermore, assume that we have a bounded loss function; specifically, suppose that $0 \le \ell((x,y);\theta) \le 1$, as in the case of the zero-one loss function. We show that uniform convergence implies generalization.

First, via telescoping sums, we can decompose the excess risk into three terms:

$$L(\hat{\theta}) - L(\theta^*) = \underbrace{L(\hat{\theta}) - \hat{L}(\hat{\theta})}_{①} + \underbrace{\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)}_{②} + \underbrace{\hat{L}(\theta^*) - L(\theta^*)}_{③}. \tag{4.3}$$

We know that $\hat{L}(\hat{\theta}) - \hat{L}(\theta^*) \leq 0$ since $\hat{\theta}$ is a minimizer of $\hat{L}$. This allows us to write

$$L(\hat{\theta}) - L(\theta^*) \leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + \hat{L}(\hat{\theta}) - \hat{L}(\theta^*) + |\hat{L}(\theta^*) - L(\theta^*)| \tag{4.4}$$

$$\leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + 0 + |\hat{L}(\theta^*) - L(\theta^*)| \tag{4.5}$$

$$\leq 2 \sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|. \tag{4.6}$$

This result tells us that if $\sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|$ is small (say, less than $\varepsilon/2$), then excess risk $L(\hat{\theta}) - L(\theta^*)$ is less than ε. But this is exactly in the form of the bound in (4.2). Hence, if we can show that a parameter family exhibits uniform convergence, we can get a bound on excess risk as well.

For future reference, Equation (4.6) can be strengthened straightforwardly into the following with slightly more careful treatment of the signs of each term:

$$L(\hat{\theta}) - L(\theta^*) \leq |\hat{L}(\theta^*) - L(\theta^*)| + L(\hat{\theta}) - \hat{L}(\hat{\theta}) \leq |\hat{L}(\theta^*) - L(\theta^*)| + \sup_{\theta \in \Theta} \Big(L(\theta) - \hat{L}(\theta) \Big) \tag{4.7}$$

This will make some of our future derivations technically slightly more convenient, but the nuanced difference between Equations (4.6) and (4.7) does not change the fundamental idea and the discussions in this chapter.

Let us try to apply our knowledge of concentration inequalities to this problem. Earlier we assumed that $\ell((x,y);\theta)$ is bounded, so we can bound 3 by $\widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$ via Hoeffding's inequality (Remark 3.4). However, we cannot apply the same concentration inequality to 1: since $\hat{\theta}$ is data-dependent by definition, the i.i.d. assumption no longer holds. (To see this, note that $\hat{\theta}$ depends on the training dataset $\{(x^{(i)}, y^{(i)})\}$, so the terms in $\widehat{L}(\hat{\theta})$, $\ell((x^{(i)}, y^{(i)}); \hat{\theta})$, all depend on the training dataset too.) This is concerning: it is certainly possible that $L(\hat{\theta}) - \hat{L}(\hat{\theta})$ is large. You've probably encountered this yourself when a model exhibits low training loss, but high validation/testing loss.

4.1.2 Deriving uniform convergence bounds

Uniform convergence is one way we can control this issue. The high-level idea is as follows:

- Suppose we have a bound of the form $\Pr[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon'] \leq \delta'$ for some single, fixed choice of θ.
- If we know *all possible values of* θ in advance, we can use the above bound to create a more general bound over all values of θ.

In particular, we can use the union-bound inequality to create the general bound described in the second bullet point, using the bound in the first bullet point:

$$\Pr\left[\forall \theta \in \Theta, |\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right] \leq \sum_{\theta \in \Theta} \Pr\left[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right]. \tag{4.8}$$

We can then use Hoeffding's inequality to deal with the summands as θ there is no longer data-dependent. We will talk more later about proving statements of this form.

4.1.3 Intuitive interpretation of uniform convergence

Since uniform convergence implies generalization, if we know that population risk and empirical risk are always "close," then excess risk is "small" as well (Figure 4.1a). In fact, it is possible to show that not only is $L(\theta)$ "close" to $\hat{L}(\theta)$ for sufficiently large data, but that the "shape" of $\hat{L}$ is "close" to the shape of L as well (Figure 4.1b). This holds for the convex case; furthermore, there are conditions under which this holds in the non-convex case, for which a rigorous treatment can be found in [Mei et al., 2017]. (*Figure design and some wording in this section were inspired by [Liang, 2016, Liu and Thomas, 2018].*)

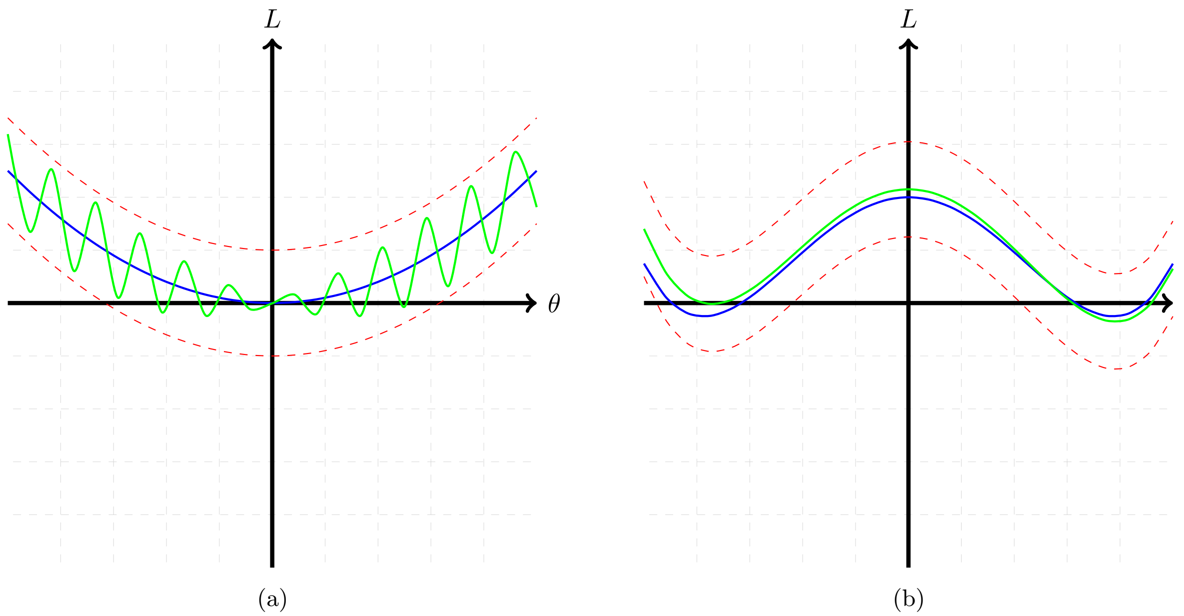

Figure 4.1: These curves demonstrate how we apply uniform convergence to bound the population risk. The blue curves are the unobserved population risk we aim to bound. The green curves denote the empirical risk we observe. Though this curve is often depicted as the fluctuating curve used in Figure 4.1a, it is more often a smooth curve whose shape mimics that of the population risk (Figure 4.1b). Uniform convergence allows us to construct additive error bounds for the excess risk, which are depicted using the red, dashed lines.

4.2 Finite hypothesis class

In this section, assume that $\mathcal{H}$ is finite. The following theorem gives a bound for the excess risk $L(\hat{h}) - L(h^*)$, where $\hat{h}$ and h^* are the minimizers of the empirical loss and population loss, respectively.

Theorem 4.1. *Suppose that our hypothesis class $\mathcal{H}$ is finite and that our loss function ℓ is bounded in $[0, 1]$, i.e. $0 \leq \ell((x, y), h) \leq 1$. Then $\forall \delta$ s.t. $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$, we have*

$$|L(h) - \hat{L}(h)| \leq \sqrt{\frac{\ln |\mathcal{H}| + \ln (2/\delta)}{2n}} \qquad \forall h \in \mathcal{H}. \tag{4.9}$$

As a corollary, we also have

$$L(\hat{h}) - L(h^*) \leq \sqrt{\frac{2(\ln |\mathcal{H}| + \ln (2/\delta))}{n}}. \tag{4.10}$$

Proof. We will prove this in two steps:

1. Use concentration inequalities to prove the bound for a fixed $h \in \mathcal{H}$, then
2. Use a union bound across the h's. (Recall that if $E_1, \ldots, E_k$ are a finite set of events, then the union bound states that $\Pr(E_1 \cup \cdots \cup E_k) \leq \sum_{i=1}^{k} \Pr(E_i)$.)

Fix some $\epsilon > 0$. By applying Hoeffding's inequality on the $\ell((x^{(i)}, y^{(i)}), h)$, we know that

$$\Pr\left(|\hat{L}(h) - L(h)| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{4.11}$$

$$= 2\exp\left(-\frac{2n^2\epsilon^2}{n}\right) \tag{4.12}$$

$$= 2\exp(-2n\epsilon^2), \tag{4.13}$$

since we can set $a_i = 0, b_i = 1$. The bound above holds for a single fixed h. To prove a similar inequality that holds for all $h \in \mathcal{H}$, we apply the union bound with $E_h = \{|\hat{L}(h) - L(h)| \geq \epsilon\}$:

$$\Pr\left(\exists h \text{ s.t. } |\hat{L}(h) - L(h)| \geq \epsilon\right) \leq \sum_{h \in \mathcal{H}} \Pr\left(|\hat{L}(h) - L(h)| \geq \epsilon\right) \tag{4.14}$$

$$\leq \sum_{h \in \mathcal{H}} 2\exp(-2n\epsilon^2) \tag{4.15}$$

$$= 2|\mathcal{H}|\exp(-2n\epsilon^2). \tag{4.16}$$

If we take δ such that $2|\mathcal{H}|\exp(-2n\epsilon^2) = \delta$, then it follows that

$$\epsilon = \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}, \tag{4.17}$$

which proves (4.9). (4.10) follows by the inequality we stated in Section 4.1.1, and taking

$$\epsilon = \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}, \tag{4.18}$$

we have that

$$\Pr\left(|L(\hat{h}) - L(h^*)| \geq \epsilon\right) \leq \Pr\left(2\sup_{h \in \mathcal{H}} |\hat{L}(h) - L(h)| \geq \epsilon\right) \tag{4.19}$$

$$\leq 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{4.20}$$

□

4.2.1 Comparing Theorem 4.1 with standard concentration inequalities

With standard concentration inequalities, we have the following bound that depends on empirical risk:

$$\forall h \in \mathcal{H}, \quad w.h.p. \quad |\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{4.21}$$

The bound here depends on each h. In contrast, the uniform convergence bound we obtain from (4.17) is uniform over all $h \in \mathcal{H}$:

$$w.h.p., \quad \forall h \in \mathcal{H}, \quad |\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{\ln|\mathcal{H}|}{\sqrt{n}}\right), \tag{4.22}$$

if we omit the $\ln(1/\delta)$ factor (we can do this since $\ln(1/\delta)$ is small in general and we take $\delta = \frac{1}{poly(n)}$). Hence, the extra $\ln|\mathcal{H}|$ term that depends on the size of our finite hypothesis family $\mathcal{H}$ can be viewed as a trade-off in order to make the bound uniform.

Remark 4.2. There is no standard definition for the term *with high probability* (*w.h.p*). For this class, the term is equivalent to the condition that the probability is higher than $1 - n^{-c}$ for some constant c.

4.2.2 Comparing Theorem 4.1 with asymptotic bounds

We can also compare the bound in Theorem 4.1 with our original asymptotic bound, namely,

$$L(\hat{h}) - L(h^*) \leq \frac{c}{n} + o\left(n^{-1}\right). \tag{4.23}$$

The $o(n^{-1})$ term can vary significantly depending on the problem. For instance, both n^{-2} and $p^{100}n^{-2}$ are $o(n^{-1})$ but the second one converges much more slowly. With the new bound, there are no longer any constants hidden in an $o(n^{-1})$ term (in fact that term is no longer there). However, we now have a slower convergence rate of $O(n^{-1/2})$.

Remark 4.3. $O(n^{-1/2})$ convergence is sometimes known as the *slow rate* while $O(n^{-1})$ convergence is known as the *fast rate*. We were only able to get the slow rate from uniform convergence: we needed asymptotics to get the fast rate. (It is possible to get the fast rate from uniform convergence under certain conditions, e.g. when the population risk on the true h^* is very low.)

4.3 Bounds for infinite hypothesis class via discretization

Unfortunately, we cannot generalize the results from the previous section directly to the case where the hypothesis class $\mathcal{H}$ is infinite, since we cannot apply the union bound to an infinite number of hypothesis functions $h \in \mathcal{H}$. However, if we consider a *bounded* and *continuous* parameterized space of $\mathcal{H}$, then we can obtain a similar uniform bound by applying a technique called *brute-force discretization.*

For this section, assume that our infinite hypothesis class $\mathcal{H}$ can be parameterized by $\theta \in \mathbb{R}^p$ with $\|\theta\|_2 \leq B$ for some fixed $B > 0$. That is, we have

$$\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq B\}. \tag{4.24}$$

The intuition behind brute-force discretization is as follows: Let $E_\theta = \{|\widehat{L}(\theta) - L(\theta)| \geq \epsilon\}$ be the "bad" events. We want the bound the probability of any one of these bad events happening (i.e. $\bigcup_\theta E_\theta$). The union bound does not work as we end up with an infinite sum. However, the union bound is very loose: these events can overlap with each other significantly. Instead, we can try to find "prototypical" bad events $E_{\theta_1}, \ldots, E_{\theta_N}$ that are somewhat disjoint so that $\bigcup_\theta E_\theta \approx \bigcup_{i=1}^N E_{\theta_i}$. We can then use the union bound on $\bigcup_{i=1}^N E_{\theta_i}$ to get a non-vacuous upper bound.

We make these ideas precise in the following section.

4.3.1 Discretization of the parameter space by ϵ-covers

We start by defining the notion of an ϵ-*cover* (also ϵ-*net*):

Definition 4.4 (ϵ-cover)**.** Let $\epsilon > 0$. An ϵ-*cover* of a set S with respect to a distance metric ρ is a subset $C \subseteq S$ such that $\forall x \in S$, $\exists x' \in C$ such that $\rho(x, x') \leq \epsilon$, or equivalently,

$$S \subseteq \bigcup_{x \in C} \mathrm{Ball}(x, \epsilon, \rho), \quad \text{where} \tag{4.25}$$

$$\mathrm{Ball}(x, \epsilon, \rho) \triangleq \{x' : \rho(x, x') \leq \epsilon\}. \tag{4.26}$$

(We note that in some definitions it is possible for points in C to lie outside of S; we do not worry about this technicality in this class.) The following lemma tells us that our parameter space $S = \{\theta \in \mathbb{R}^p : \|\theta\|_2 \leq B\}$ has an ϵ-cover with not too many elements:

Lemma 4.5 (ϵ-cover of ℓ_2 ball)**.** *Let $B, \epsilon > 0$, and let $S = \{x \in \mathbb{R}^p : \|x\|_2 \leq B\}$. Then there exists an ϵ-cover of S with respect to the ℓ_2-norm with at most* $\max\left(\left(\frac{3B\sqrt{p}}{\epsilon}\right)^p, 1\right)$ *elements.*

Proof. Note that if $\epsilon > B\sqrt{p}$, then S is trivially contained in the ball centered at the origin with radius ϵ and the ϵ-cover has size 1. Assume $\epsilon \le B\sqrt{p}$. Set

$$C = \left\{ x \in S : x_i = k_i \frac{\epsilon}{\sqrt{p}}, k_i \in \mathbb{Z}, |k_i| \le \frac{B\sqrt{p}}{\epsilon} \right\}, \tag{4.27}$$

i.e. C is the set of grid points in $\mathbb{R}^p$ of width $\frac{\epsilon}{\sqrt{p}}$ that are contained in S. See Figure 4.2 for an illustration.

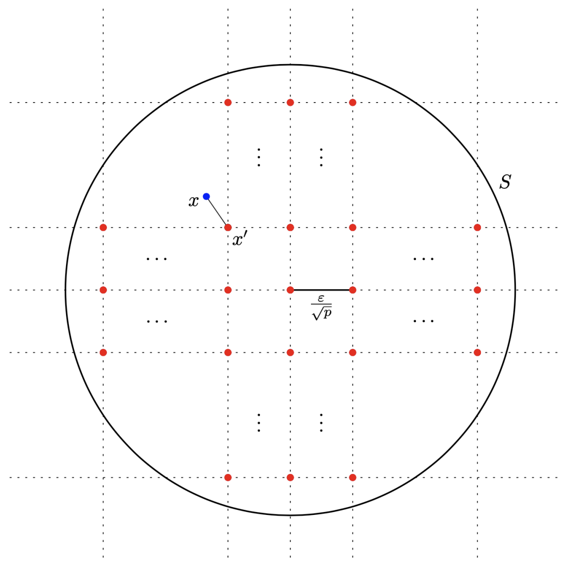

Figure 4.2: The ϵ-cover (shown in red) of S that we construct in the proof of Lemma 4.5. For $x \in S$, we choose the grid point x' such that $\|x - x'\|_2 \le \epsilon$.

We claim that C is an ϵ-cover of S with respect to the ℓ_2-norm: $\forall x \in S$, there exists a grid point $x' \in C$ such that $|x_i - x'_i| \le \frac{\epsilon}{\sqrt{p}}$ for each i. Therefore,

$$\|x - x'\|_2 = \sqrt{\sum_{i=1}^{p} |x_i - x'_i|^2} \le \sqrt{p \cdot \frac{\epsilon^2}{p}} = \epsilon.$$

We now bound the size of C. Since each k_i in the definition of C has at most $2\frac{B\sqrt{p}}{\epsilon} + 1$ choices, we have

$$|C| \le \left(\frac{2B\sqrt{p}}{\epsilon} + 1 \right)^p \le \left(\frac{3B\sqrt{p}}{\epsilon} \right)^p. \tag{4.28}$$

□

Remark 4.6. We can actually prove a stronger version of Lemma 4.5: there exists an ϵ-cover of S with at most $\left(\frac{3B}{\epsilon}\right)^p$ elements. We will be using this version of the lemma in the proof below. (We will leave the proof of this stronger version as a homework exercise.)

4.3.2 Uniform convergence bound for infinite $\mathcal{H}$

Definition 4.7 (κ-Lipschitz functions)**.** Let $\kappa \ge 0$ and $\|\cdot\|$ be a norm on the domain D. A function $L : D \to \mathbb{R}$ is said to be *κ-Lipschitz* with respect to $\|\cdot\|$ if for all $\theta, \theta' \in D$, we have

$$|L(\theta) - L(\theta')| \le \kappa \|\theta - \theta'\|.$$

Assume that our infinite hypothesis class $\mathcal{H}$ can be parameterized by $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq B\}$. We have the following uniform convergence theorem for our infinite hypothesis class $\mathcal{H}$:

Theorem 4.8. *Suppose $\ell((x, y), \theta) \in [0, 1]$, and $\ell((x, y), \theta)$ is κ-Lipschitz in θ with respect to the ℓ_2-norm for all (x, y). Then, with probability at least $1 - O(\exp(-\Omega(p)))$, we have*

$$\forall \theta, \quad |\hat{L}(\theta) - L(\theta)| \leq O\left(\sqrt{\frac{p \max(\ln(\kappa B n), 1)}{n}}\right). \tag{4.29}$$

Proof of Theorem 4.8. Fix parameters $\delta, \epsilon > 0$ (we will specify their values later). Let C be the ϵ-cover of our parameter space S with respect to the ℓ_2-norm constructed in Lemma 4.5. Define event $E = \left\{\forall \theta \in C,\ |\hat{L}(\theta) - L(\theta)| \leq \delta\right\}$. By Theorem 4.1, we have $\Pr(E) \geq 1 - 2|C| \exp(-2n\delta^2)$.

Now for any $\theta \in S$, we can pick some $\theta_0 \in C$ such that $\|\theta - \theta_0\|_2 \leq \epsilon$. Since L and $\widehat{L}$ are κ-Lipschitz functions (this follows from the Lipschitzness of ℓ), we have

$$|L(\theta) - L(\theta_0)| \leq \kappa\|\theta - \theta_0\|_2 \leq \kappa\epsilon, \text{ and} \tag{4.30}$$

$$|\hat{L}(\theta) - \hat{L}(\theta_0)| \leq \kappa\|\theta - \theta_0\|_2 \leq \kappa\epsilon. \tag{4.31}$$

Therefore, conditional on E, we have

$$|\hat{L}(\theta) - L(\theta)| \leq |\hat{L}(\theta) - \hat{L}(\theta_0)| + |\hat{L}(\theta_0) - L(\theta_0)| + |L(\theta_0) - L(\theta)| \leq 2\kappa\epsilon + \delta. \tag{4.32}$$

It remains to choose suitable parameters δ and ϵ to get the desired bound in Theorem 4.8 while making the failure probability small. First, set $\epsilon = \delta/(2\kappa)$ so that conditional on E,

$$|\hat{L}(\theta) - L(\theta)| \leq 2\delta. \tag{4.33}$$

To choose the correct δ, we must reason about the probability of E under different choices of the parameter. The event E happens with probability $1 - 2|C| \exp(-2n\delta^2) = 1 - 2\exp(\ln|C| - 2n\delta^2)$. From Remark 4.6, we know that $\ln|C| \leq p \ln(3B/(\delta/2))$. If we ignore the log term and assume $\ln|c| \leq p$, then this would give us the high probability bound we want:

$$2|C| \exp(-2n\delta^2) = 2\exp(\ln|C| - 2n\delta^2) \leq 2\exp(p - 2p) = 2\exp(-p). \tag{4.34}$$

(At the same time, we see from (4.33) that this choice of δ gives $|\hat{L}(\theta) - L(\theta)| \leq 2\sqrt{\frac{p}{n}}$, which is roughly the bound we want.)

Since we cannot actually drop the log term in the inequality $\ln|C| \leq p \ln(3B/(\delta/2))$, we need to make δ a little bit bigger. So, if we set $\delta = \sqrt{\frac{c_0 p \max(1, \ln(\kappa B n))}{n}}$ with $c_0 = 36$, then by Remark 4.6,

$$\begin{aligned}
\ln|C| - 2n\delta^2 &\leq p \ln\left(\frac{6B\kappa}{\delta}\right) - 2n\delta^2 && (4.35)\\
&\leq p \ln\left(\frac{6B\kappa\sqrt{n}}{\sqrt{c_0 p \max(1, \ln(\kappa B n))}}\right) - 2n\frac{c_0 p}{n}\ln(\kappa B n) && \text{(dfn of } \delta) \quad (4.36)\\
&\leq p \ln\left(\frac{B\kappa\sqrt{n}}{\sqrt{p}}\right) - 72p\ln(\kappa B n) && (\max(1, \ln(\kappa B n)) \geq 1, c_0 = 36) \quad (4.37)\\
&\leq p\ln(B\kappa n) - 72p\ln(B\kappa n) && (\sqrt{n/p} \leq n) \quad (4.38)\\
&\leq -p, && (4.39)
\end{aligned}$$

since $\ln(B\kappa n) \geq 1$ for large enough n. Therefore, with probability greater than $1 - 2|C| \exp(-2n\delta^2) = 1 - 2\exp(\ln|C| - 2n\delta^2) \geq 1 - O(e^{-p})$, we have

$$|\hat{L}(\theta) - L(\theta)| \leq 2\delta = O\left(\sqrt{\frac{p}{n}\max(1, \ln(\kappa B n))}\right). \tag{4.40}$$

□

Remark 4.9. We bounded the generalization error $|\hat{L}(\theta) - L(\theta)|$ by $\delta + 2\epsilon\kappa \le \sqrt{\frac{\ln|C|}{n}} + 2\epsilon\kappa$. The term $2\epsilon\kappa$ represents the error from our brute-force discretization. It is not a problem because we can always choose ϵ small enough without worrying about the growth of the first term $\sqrt{\frac{\ln|C|}{n}}$. This in turn is because $\ln|C| \approx p \ln \epsilon^{-1}$, which is very insensitive to ϵ, even if we let $\epsilon = \frac{1}{poly(n)}$. We also observe that both $\sqrt{\frac{\ln|C|}{n}}$ and $\sqrt{\frac{p}{n}}$ are bounds that depend on the "size" of our hypothesis class, in terms of either its total size or dimensionality. This possibly explains why one may need more training samples when the hypothesis class is larger.

4.4 Rademacher complexity

4.4.1 Motivation for a new complexity measure

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. We previously 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 sections, we derived bounds for the generalization gap in two cases:

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

$$L(\hat{h}) - \hat{L}(\hat{h}) \le \tilde{O}\left(\sqrt{\frac{\log|\mathcal{H}|}{n}}\right). \tag{4.41}$$

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

$$L(\hat{h}) - \hat{L}(\hat{h}) \le \tilde{O}\left(\sqrt{\frac{p}{n}}\right). \tag{4.42}$$

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 measures of $\mathcal{H}$.

The bound (4.42) 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:

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

(4.42) 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 to prove a better bound.

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

$$L(\hat{h}) - \hat{L}(\hat{h}) \le \tilde{O}\left(\sqrt{\frac{\text{Complexity}(\Theta)}{n}}\right). \tag{4.43}$$

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.

4.4.2 Definitions

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.

$$\mathbb{E}\left[\sup_{h\in H}(L(h)-\hat{L}(h))\right] \leq \text{ upper bound.} \tag{4.44}$$

The expectation is over the randomness in the training data $\{(x^{(i)},y^{(i)})\}_{i=1}^n$.[1]

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

Definition 4.10 (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

$$R_n(F) \triangleq \mathop{\mathbb{E}}_{z_1,\dots,z_n \overset{\text{iid}}{\sim} P}\left[\mathop{\mathbb{E}}_{\sigma_1,\dots,\sigma_n \overset{\text{iid}}{\sim}\{\pm 1\}}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right]\right], \tag{4.45}$$

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

Remark 4.11. For applications to empirical risk minimization, we will take $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$. However, Definition 4.10 holds for abstract input spaces $\mathcal{Z}$ as well.

Remark 4.12. 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 that $R_n(\mathcal{F})$ is the maximal possible correlation between outputs of some $f\in\mathcal{F}$ (on points $f(z_1),\dots,f(z_n)$) and random Rademacher variables $(\sigma_1,\dots,\sigma_n)$. 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 4.13.

$$\mathop{\mathbb{E}}_{z_1,\dots,z_n \overset{\text{iid}}{\sim} P}\left[\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathop{\mathbb{E}}_{z\sim P}[f(z)]\right]\right] \leq 2R_n(\mathcal{F}). \tag{4.46}$$

Remark 4.14. We can think of $\frac{1}{n}\sum_{i=1}^n f(z_i)$ as an empirical average and $\mathbb{E}_{z\sim P}[f(z)]$ as a population average.

Why is Theorem 4.13 useful to us? We can set $\mathcal{F}$ to be the family of loss functions, i.e.

$$\mathcal{F} = \{z=(x,y)\in\mathcal{Z}\mapsto \ell((x,y),h)\in\mathbb{R} : h\in\mathcal{H}\}. \tag{4.47}$$

This is the family of losses induced by the hypothesis functions in $\mathcal{H}$. We also define the function class $-\mathcal{F}$ as $\{-f : f\in\mathcal{F}\}$. It should be obvious from this definition that $R_n(\mathcal{F}) = R_n(-\mathcal{F})$ since $\sigma_i \overset{d}{=} -\sigma_i$ for all i. Then, letting $z_i = (x^{(i)},y^{(i)})$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(L(h)-\hat{L}(h)\right)\right] &= \mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\}}\left[\sup_{h\in\mathcal{H}}\left[L(h)-\frac{1}{n}\sum_{i=1}^n \ell((x^{(i)},y^{(i)}),h)\right]\right] && (4.48)\\
&= \mathop{\mathbb{E}}_{\{z_i\}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right)\right] && (4.49)\\
&= \mathop{\mathbb{E}}_{\{z_i\}}\left[\sup_{f\in-\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f(z)]\right)\right] && (4.50)\\
&\leq 2R_n(-\mathcal{F}) = 2R_n(\mathcal{F}) && (4.51)
\end{aligned}$$

[1] Though we might like to pull the sup outside of the $\mathbb{E}$ operator, and bound the expectation of the excess risk (a far simpler quantity to deal with!), in general, the sup and $\mathbb{E}$ operators do not commute. In particular, $\mathbb{E}\left[\sup_{h\in\mathcal{H}}(L(h)-\hat{L}(h))\right] \geq \sup_{h\in\mathcal{H}}\mathbb{E}\left[L(h)-\hat{L}(h)\right]$.

where the last step follows by Theorem 4.13.

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 $\ell((x^{(1)}, y^{(1)}), h), \dots \ell((x^{(n)}, y^{(n)}), h)$ correlates with $\sigma_1, \dots, \sigma_n$.

Example 4.15. Consider the binary classification setting where $y \in \{\pm 1\}$. Let ℓ_{0-1} denote the zero-one loss function. Note that

$$\ell_{0-1}((x,y),h) = \mathbf{1}\{h(x) \neq y\} = \frac{1 - yh(x)}{2}. \tag{4.52}$$

Hence,

$$\begin{aligned}
R_n(\mathcal{F}) &= \mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\},\sigma_i}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell_{0-1}((x^{(i)},y^{(i)}),h)\sigma_i\right] && \text{(by definition)} && (4.53)\\
&= \mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\},\sigma_i}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \left(\frac{-h(x^{(i)})y^{(i)}+1}{2}\right)\sigma_i\right] && \text{(by (4.52))} && (4.54)\\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\},\sigma_i}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i + \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n -h(x^{(i)})y^{(i)}\sigma_i\right] && (\sup \text{only over } \mathcal{H}) && (4.55)\\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\},\sigma_i}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n -h(x^{(i)})y^{(i)}\sigma_i\right] && (\mathbb{E}[\sigma_i]=0) && (4.56)\\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\{(x^{(i)},y^{(i)})\},\sigma_i}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n h(x^{(i)})\sigma_i\right] && (-y_i\sigma_i \overset{d}{=} \sigma_i) && (4.57)\\
&= \frac{1}{2}R_n(\mathcal{H}). && \text{(by definition)} && (4.58)
\end{aligned}$$

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 4.16. Rademacher complexity is invariant to translation. This property manifests in the previous example when 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 4.13.

Proof of Theorem 4.13. We use a technique called *symmetrization*, which is a very important technique in probability theory. We first fix $z_1, \dots, z_n$and draw $z_1', \dots z_n' \overset{\text{iid}}{\sim} P$. Then we can rewrite the term in the expectation on the LHS of (4.46):

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f]\right) &= \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathop{\mathbb{E}}_{z_1',\dots,z_n'}\left[\frac{1}{n}\sum_{i=1}^n f(z_i')\right]\right) && (4.59)\\
&= \sup_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{z_1',\dots,z_n'}\left[\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(z_i')\right]\right) && (4.60)\\
&\leq \mathop{\mathbb{E}}_{z_1',\dots,z_n'}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(z_i')\right)\right]. && (4.61)
\end{aligned}$$

The last inequality is because in general,

$$\sup_u\left(\mathop{\mathbb{E}}_v[g(u,v)]\right) \leq \sup_u\left(\mathop{\mathbb{E}}_v\left[\sup_{u'}(g(u',v))\right]\right) = \mathop{\mathbb{E}}_v\left[\sup_u(g(u,v))\right] \tag{4.62}$$

since the sup over u becomes vacuous after we replace u with u'.

Now, if we take the expectation over $z_1, \ldots, z_n$ for both sides of (4.61),

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}[f]\right)\right] \le \mathop{\mathbb{E}}_{z_i}\left[\mathop{\mathbb{E}}_{z_i'}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n (f(z_i) - f(z_i'))\right)\right]\right] \tag{4.63}$$

$$= \mathop{\mathbb{E}}_{z_i, z_i'}\left[\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(z_i) - f(z_i')\right)\right)\right]\right] \tag{4.64}$$

$$\le \mathop{\mathbb{E}}_{z_i, z_i', \sigma_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right) + \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n -\sigma_i f(z_i')\right)\right] \tag{4.65}$$

$$= 2R_n(\mathcal{F}), \tag{4.66}$$

where (4.64) 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 4.13 achieves. Consider the quantities on the LHS and RHS of (4.46):

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

First, we removed $\mathbb{E}[f(z)]$, 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.

4.4.3 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 $-1 \le f(z_0) \le 1$ for all $f \in \mathcal{F}$. Then

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim P}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right] = \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}f(z_0)\sum_{i=1}^n \sigma_i\right] \tag{4.67}$$

$$\le \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\right|\right] \qquad \text{(since } f(z_0) \in [-1,1]) \tag{4.68}$$

$$\le \mathop{\mathbb{E}}_{\sigma_i}\left[\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2\right]^{\frac{1}{2}} \qquad \text{(Jensen's Inequality)} \tag{4.69}$$

$$= \frac{1}{n}\left(\mathop{\mathbb{E}}_{\sigma_i,\sigma_j}\left[\sum_{i,j=1}^n \sigma_i\sigma_j\right]\right)^{\frac{1}{2}} \tag{4.70}$$

$$= \frac{1}{n}\left(\mathop{\mathbb{E}}_{\sigma_i}\left[\sum_{i=1}^n \sigma_i^2\right]\right)^{\frac{1}{2}} \tag{4.71}$$

$$= \frac{1}{n}\cdot\sqrt{n} = \frac{1}{\sqrt{n}}. \tag{4.72}$$

This bound does not depend on $\mathcal{F}$ (except on the fact that $f \in \mathcal{F}$ is bounded). This example illustrates that a bound on the Rademacher complexity can sometimes depend only on the (known) distribution of the Rademacher random variables.