이전 section에서 한 것: bound 기댓값 of sup[.]
=> training example z_1, ..., z_n에 대한 기댓값임.
=> 그러나 주로 only ONE TRAINING SET이어서, 그렇게까지 많은 training set이 없음
=> 즉, 이 bound는 우리의 ONE TRAINING SET이 bound됨을 보장해주지 못함
=> 따라서 이 값 자체 (즉 sup[.])를 높은 확률로 bound해주는 걸 할거임!

4.5 Empirical Rademacher complexity

In the previous section, we bounded the expectation of $\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z\sim P}[f(z)]\right]$. 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 4.17 (Empirical Rademacher complexity)**.** Given a dataset $S = \{z_1, \ldots, z_n\}$, the *empirical Rademacher complexity* is defined as

$$R_S(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]. \tag{4.73}$$

<= 얘의 expectation이 Rademacher complexity임.

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

As the name suggests, the expectation of the empirical Rademacher complexity is the Rademacher complexity:

$$R_n(\mathcal{F}) = \mathop{\mathbb{E}}_{\substack{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P \\ S=\{z_1,\ldots,z_n\}}}[R_S(\mathcal{F})]. \tag{4.74}$$

Theorem 4.18. *Suppose for all $f \in \mathcal{F}$, $0 \le f(z) \le 1$. Then, with probability at least $1-\delta$,*

원래는 이 LHS의 기댓값을 bound했었는데, (on the training examples z_1 to z_n) 이제는 LHS 자체를 bound하는거

$$\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f(z)]\right] \le 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4.75}$$

Proof. For conciseness, define

define the LHS as....

$$g(z_1,\ldots,z_n) \triangleq \sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f(z)]\right]. \tag{4.76}$$

We prove the theorem in 4 steps.

McDiarmid: Section 3.5, Theorem 3.15, page 20
함수 f가 bounded difference condition (변수 하나가 바뀌어도 함수 값이 많이 움직이지 않음)을 만족하면, any indep RVs X_1, ..., X_n에 대해, Pr[f(X_1, ... , X_n)- E[f(X_1, ..., X_n)] >= t] <= exp(-2t^2 / sigma c_i ^2)

<STEP 1> McDiarmid를 통해 g를 bound함.

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

(4.77)~(4.80): 함수 g의 bounded difference condition 확인

$$g(z_1,\ldots,z_n) - g(z_1,\ldots,z_i',\ldots,z_n) \le \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{j=1}^{n} f(z_j)\right] - \sup_{f\in\mathcal{F}}\left[\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} f(z_j)\right) + \frac{f(z_i')}{n}\right] \tag{4.77}$$

$$\le \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{j=1}^{n} f(z_j) - \left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} f(z_j)\right) - \frac{f(z_i')}{n}\right] \tag{4.78}$$

$$= \sup_{f\in\mathcal{F}}\left[\frac{1}{n}\left(f(z_i) - f(z_i')\right)\right] \tag{4.79}$$

$$\le \frac{1}{n}. \tag{4.80}$$

(4.78) holds because in general, $\sup_f A(f) - \sup_f B(f) \le \sup_f [A(f) - B(f)]$, and (4.80) holds since f is bounded by $[0,1]$. We can thus apply McDiarmid's Inequality with parameters $c_1 = \cdots = c_n = 1/n$:

$$\Pr\left[g(z_1,\ldots,z_n) \ge \mathop{\mathbb{E}}_{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P}[g] + \epsilon\right] \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp(-2n\epsilon^2). \tag{4.81}$$

Step 2: We apply Theorem 4.13 to get

$$\mathbb{E}_{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P}[g] \leq 2R_n(\mathcal{F}). \tag{4.82}$$

Step 3: Define

$$\tilde{g}(z_1,\ldots,z_n) = R_S(\mathcal{F}) \triangleq \mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]. \tag{4.83}$$

Using a similar argument to that of Step 1, we show that $\tilde{g}$ satisfies the "bounded difference condition."

$$\begin{aligned}
&\tilde{g}(z_1,\ldots,z_n) - \tilde{g}(z_1,\ldots,z_i',\ldots,z_n) \\
&\leq \mathbb{E}_{\sigma_i}\left[\sup_{f\in F}\left[\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j)\right] - \sup_{f\in F}\left[\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n}\sigma_j f(z_j)\right) + \frac{1}{n}\sigma_i f(z_i')\right]\right] &(4.84)\\
&\leq \mathbb{E}_{\sigma_i}\left[\sup_{f\in F}\left(\frac{1}{n}\sigma_i(f(z_i) - f(z_i'))\right)\right] &(4.85)\\
&\leq \frac{1}{n}, &(4.86)
\end{aligned}$$

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

$$\Pr[\tilde{g} - \mathbb{E}[\tilde{g}] \geq \epsilon] \leq \exp(-2n\epsilon^2), \quad \text{and} \quad \Pr[\tilde{g} - \mathbb{E}[\tilde{g}] \leq -\epsilon] \leq \exp(-2n\epsilon^2). \tag{4.87}$$

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 $\geq 1-\delta$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f]\right] &= g \leq \mathbb{E}[g] + \epsilon &\text{(Step 1)} &(4.88)\\
&\leq 2R_n(\mathcal{F}) + \epsilon &\text{(Step 2)} &(4.89)\\
&\leq 2(R_S(\mathcal{F}) + \epsilon) + \epsilon &\text{(Step 3)} &(4.90)\\
&= 2R_S(\mathcal{F}) + 3\epsilon, &&(4.91)
\end{aligned}$$

as required. □

Setting $\mathcal{F}$ to be a family of loss functions bounded by $[0,1]$ in Theorem 4.18 gives the following corollary:

Corollary 4.19. Let $\mathcal{F}$ 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

$$\hat{L}(h) - L(h) \leq 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{for all } h \in \mathcal{H}. \tag{4.92}$$

Remark 4.20. If we want to bound the generalization gap by the average Rademacher complexity instead, we can replace the RHS of (4.92) with $2R_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}$.

Interpretation of Corollary 4.19. It is typically the case that $O\left(\sqrt{\frac{\log(2/\delta)}{n}}\right) \ll R_S(\mathcal{F})$ and $O\left(\sqrt{\frac{\log(2/\delta)}{n}}\right) \ll R_n(\mathcal{F})$. This is the case because $R_S(\mathcal{F})$ and $R_n(\mathcal{F})$ often take the form $\frac{c}{\sqrt{n}}$ where c

is a big constant depending on the complexity of $\mathcal{F}$, whereas we only have a logarithmic term in the numerator of $O\left(\sqrt{\frac{\log(2/\delta)}{n}}\right)$. As a result, we can view the $3\sqrt{\frac{\log(2/\delta)}{n}}$ term in the RHS of Corollary 4.19 as negligible. Another way of seeing this is noting that a $\widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$ term is necessary even for the concentration bound of a single function $h \in \mathcal{H}$. Previously, we bounded $L(h) - \hat{L}(h)$ using a union bound over $h \in \mathcal{H}$, which necessarily needs to be larger than $\widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$. As a result, the $O\left(\sqrt{\frac{\log(2/\delta)}{n}}\right)$ term is not significant.

4.5.1 Rademacher complexity is translation invariant

A useful fact is that both empirical Rademacher complexity and average Rademacher complexity are translation invariant. (This is not obvious when thinking of how translation affects the picture in Figure 4.3.)

Proposition 4.5.1. Let $\mathcal{F}$ be a family of functions mapping $Z \mapsto \mathbb{R}$ and define $\mathcal{F}' = \{f'(z) = f(z) + c_0 \mid f \in \mathcal{F}\}$ for some $c_0 \in \mathbb{R}$. Then $R_S(\mathcal{F}) = R_S(\mathcal{F}')$ and $R_n(\mathcal{F}) = R_n(\mathcal{F}')$.

Proof. We will prove here that empirical Rademacher complexity is translation invariant.

$$R_S(\mathcal{F}') = \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f' \in \mathcal{F}'} \frac{1}{n}\sum_{i=1}^n \sigma_i f'(z_i)\right] \tag{4.93}$$

$$= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i (f(z_i) + c_0)\right] \tag{4.94}$$

$$= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i c_0 + \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right] \tag{4.95}$$

$$= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right] = R_S(\mathcal{F}), \tag{4.96}$$

where (4.96) follows because $\mathbb{E}_{\sigma_1,\ldots,\sigma_n} \frac{1}{n}\sum_{i=1}^n \sigma_i c_0 = 0$, since the σ_i's are Rademacher random variables. □

4.6 Covering number upper bounds Rademacher complexity

In Chapter 5, we will prove Rademacher complexity bounds that hinge on elegant, ad-hoc algebraic manipulations that may not extend to more general settings. Here, we consider a more fundamental approach for proving empirical Rademacher complexity bounds based on coverings of the output space. The trade-off is generally more tedium.

The first important observation is that for purposes of computing the **empirical** Rademacher complexity on samples $z_1, \ldots, z_n$,

$$R_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right], \tag{4.97}$$

we only care about the "output" of function $f \in \mathcal{F}$, and not the function itself (i.e. it is sufficient for our purposes to know $f(z_1), \ldots, f(z_n)$, but not know f). In other words, we can characterize $f \in \mathcal{F}$ by $f(z_1), \ldots, f(z_n)$. In the sequel, we will take advantage of this simplification from the (potentially large) space of all functions $\mathcal{F}$ to the *output space*,

$$\mathcal{Q} \triangleq \left\{\big(f(z_1), \ldots, f(z_n)\big)^\top : f \in \mathcal{F}\right\} \subseteq \mathbb{R}^n, \tag{4.98}$$

which may be drastically smaller than $\mathcal{F}$. Correspondingly, the empirical Rademacher complexity can be rewritten as a maximization over the output space Q instead of the function space $\mathcal{F}$:

$$R_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{v\in Q}\frac{1}{n}\langle\sigma, v\rangle\right]. \tag{4.99}$$

In other words, the "complexity of $\mathcal{F}$" can be also interpreted as "how much the vectors in Q can be correlated with a random vector σ." See Figure 4.3 for an illustration of this idea. One can also view $\mathbb{E}_\sigma\left[\sup_{v\in Q}\frac{1}{n}\langle\sigma, v\rangle\right]$ as a complexity measure for the set Q. If we replace σ by a Gaussian vector with spherical covariance, then the corresponding quantity (without the $\frac{1}{n}$ scaling), $\mathbb{E}_{g\sim N(0,I)}\left[\sup_{v\in Q}\langle g, v\rangle\right]$, is often referred to as the Gaussian complexity of the set Q. (It turns out that Gaussian complexity and Rademacher complexity are closely related.)

Another corollary of this is that the empirical Rademacher complexity only depends on the functionality of $\mathcal{F}$ but not on the exact parameterization of $\mathcal{F}$. For example, suppose we have two parameterizations $\mathcal{F} = \{f(x) = \sum\theta_i x_i \mid \theta\in\mathbb{R}^d\}$ and $\mathcal{F}' = \{f(x) = \sum\theta_i^3\cdot w_i x_i \mid \theta\in\mathbb{R}^d, w\in\mathbb{R}^d\}$. Since $Q_\mathcal{F}$ and $Q_{\mathcal{F}'}$ are the same, we see that $R_S(\mathcal{F}) = R_S(\mathcal{F}')$ since our earlier expression for $R_S(\mathcal{F})$ only depends on $\mathcal{F}$ through $Q_\mathcal{F}$.

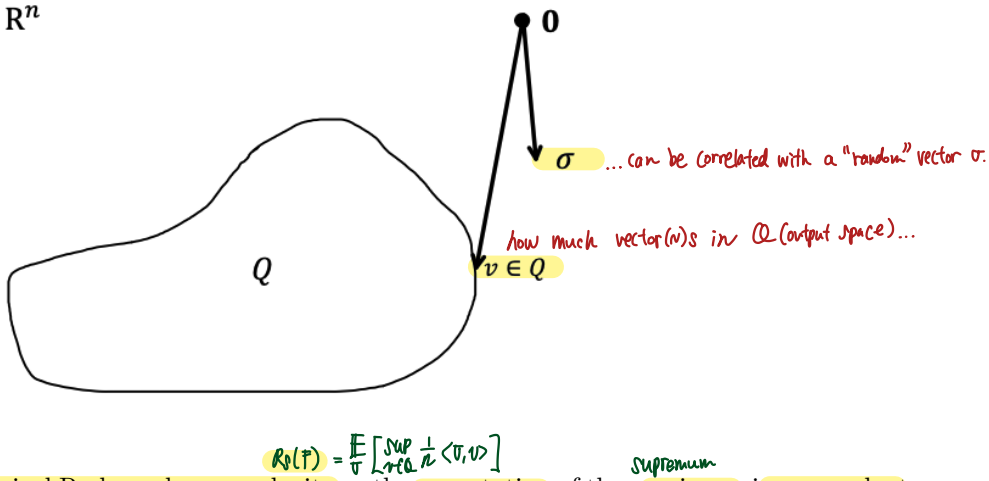

Figure 4.3: We can view empirical Rademacher complexity as the expectation of the maximum inner product between σ and $v \in Q$.

Rademacher complexity of finite hypothesis classes. In practice, we cannot directly evaluate the Rademacher complexity, so we instead bound its value using quantities that are computable. Given finite $|\mathcal{Q}|$, we often rely on the following bound, which is also known as Massart's finite lemma:

Proposition 4.6.1. Let $\mathcal{F}$ be a collection of functions mapping $Z \mapsto \mathbb{R}$ and let $\mathcal{Q}$ be defined as in (4.98). Assume that $\frac{1}{\sqrt{n}}\|v\|_2 \le M < \infty$ for all $v \in \mathcal{Q}$. Then,

$$R_S(\mathcal{F}) \le \sqrt{\frac{2M^2\log|\mathcal{Q}|}{n}} \tag{4.100}$$

We prove a (slightly) simplified version of this result in Problem 3(c) of Homework 2, so we omit the proof of Massart's lemma here. Using Massart's lemma, we can also bound the Rademacher complexity in terms of $\mathcal{F}$. Restating the assumption accordingly,

Corollary 4.21. Let $\mathcal{F}$ be a collection of functions mapping $Z \mapsto \mathbb{R}$. If $\sqrt{\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2} \leq M$ for all $f \in \mathcal{F}$, then

$$R_S(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log |\mathcal{F}|}{n}}. \tag{4.101}$$

Note that Corollary 4.21 yields a looser bound than Massart's lemma since $|\mathcal{Q}| \leq |\mathcal{F}|$.

In practice, we rarely apply Massart's lemma directly since $|\mathcal{Q}|$ is typically infinite. In the sequel, we discuss alternative approaches to bounding the Rademacher complexity that are appropriate for this setting.

Bounding "Rademacher complexity" using ϵ-covers. When $|\mathcal{Q}|$ is infinite, we can apply the same discretization trick that we used to prove the generalization bound for an infinite-hypothesis space. This time, instead of trying to cover the parameter space, we will cover the output space. To this end, we first recall a few definitions concerning ϵ-covers.

Definition 4.22. $\mathcal{C}$ is an *ϵ-cover* of $\mathcal{Q}$ with respect to metric ρ if for all $v' \in \mathcal{Q}$, there exists $v \in \mathcal{C}$ such that $\rho(v, v') \leq \epsilon$.

Definition 4.23. The *covering number* is defined as the minimum size of an ϵ-cover, or explicitly:

$$N(\epsilon, \mathcal{Q}, \rho) \triangleq (\text{min size of } \epsilon\text{-cover of } \mathcal{Q} \text{ w.r.t. metric } \rho). \tag{4.102}$$

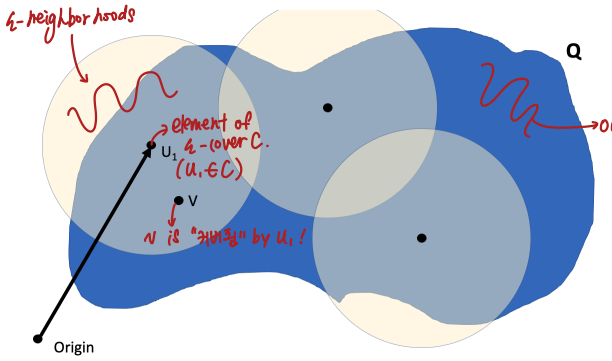

Figure 4.4: We can visualize the ϵ-cover $\mathcal{C}$ by depicting a set of ϵ-balls that cover the output space $\mathcal{Q}$. The yellow circles denote the ϵ-neighborhoods of the covering points $u_i \in \mathcal{C}$.

In subsequent derivations, we will use the metric $\rho(v, v') = \frac{1}{\sqrt{n}}\|v - v'\|_2$.

Remark 4.24. We normalize the ℓ_2 norm (in ρ) by $\frac{1}{\sqrt{n}}$ to simplify comparisons to the functional analysis view of the Rademacher complexity. In the literature, the ϵ-cover of $\mathcal{Q}$ defined above is also referred to as an ϵ-cover of the function class $\mathcal{F}$ under the $L_2(P_n)$ metric.[2] In particular,

$$L_2(P_n)(f, f') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - f'(z_i))^2}. \tag{4.103}$$

[2] P_n denotes the empirical distribution, i.e. the uniform distribution over the observations $z_1, \ldots, z_n$. More generally the $L_p(Q)$ metric is defined by $\left[\mathbb{E}_Q\left[(f(z) - f'(z))^p\right]\right]^{1/p}$.

Recall we have established the following correspondences between the set of functions $\mathcal{F}$ and the output space $\mathcal{Q}$:

$$f \in \mathcal{F} \iff \begin{pmatrix} f(z_1) \\ \vdots \\ f(z_n) \end{pmatrix} \in \mathcal{Q} \tag{4.104}$$

We can write a trivial correspondence between both the output and function class points of view as follows:

$$N(\epsilon, \mathcal{F}, L_2(P_n)) = N\left(\epsilon, \mathcal{Q}, \frac{1}{\sqrt{n}}\|\cdot\|_2\right) \tag{4.105}$$

The results below will be stated in the function-space notation, but in the proofs we will shift to the $\mathcal{Q}$-formulation for the sake of clarity. In general, we prefer to reason about covering numbers on $\mathcal{Q}$ as it is more natural to analyze vector spaces (compared to function spaces).

Equipped with the definition of minimal ϵ-covers, we can prove the following Rademacher complexity bound:

Theorem 4.25. *Let $\mathcal{F}$ be a family of functions $Z \mapsto [-1, 1]$. Then*

$$R_S(\mathcal{F}) \leq \inf_{\epsilon > 0}\left(\epsilon + \sqrt{\frac{2\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}}\right). \tag{4.106}$$

The ϵ term can be thought of as the discretization error, while the second term is the Rademacher complexity of the finite ϵ-cover. The precise form of this complexity bound follows from Proposition 4.6.1.

Proof. Fix any $\epsilon > 0$. Let $\mathcal{C}$ be the minimal ϵ-cover of $\mathcal{Q}$ with respect to the metric $\rho(v, v') = \frac{1}{\sqrt{n}}\|v - v'\|_2$. Note that $|\mathcal{C}| = N(\epsilon, \mathcal{Q}, \frac{1}{\sqrt{n}}\|\cdot\|_2) = N(\epsilon, \mathcal{F}, L_2(P_n))$.

We aim to bound $R_S(\mathcal{F}) = \mathbb{E}_\sigma[\sup_{v\in\mathcal{Q}} \frac{1}{n}\langle v, \sigma\rangle]$ by approximating v with $v' \in \mathcal{C}$. In particular, for every point $v \in \mathcal{Q}$, choose $v' \in \mathcal{C}$ such that $\rho(v, v') \leq \epsilon$ and z is small (specifically, $\frac{1}{\sqrt{n}}\|z\|_2 \leq \epsilon$). This gives

$$\frac{1}{n}\langle v, \sigma\rangle = \frac{1}{n}\langle v', \sigma\rangle + \frac{1}{n}\langle v - v', \sigma\rangle \tag{4.107}$$

$$\leq \frac{1}{n}\langle v', \sigma\rangle + \frac{1}{n}\|z\|_2\|\sigma\|_2 \qquad (z = v - v', \text{ Cauchy-Schwarz}) \tag{4.108}$$

$$\leq \frac{1}{n}\langle v', \sigma\rangle + \epsilon. \qquad (\text{since } \|z\|_2 \leq \sqrt{n}\epsilon \text{ and } \|\sigma\|_2 \leq \sqrt{n}) \tag{4.109}$$

Taking the expectation of the supremum on both sides of this inequality gives

$$R_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{v\in\mathcal{Q}} \frac{1}{n}\langle v, \sigma\rangle\right] \tag{4.110}$$

$$\leq \mathbb{E}_\sigma\left[\sup_{v'\in\mathcal{C}}\left(\frac{1}{n}\langle v', \sigma\rangle + \epsilon\right)\right] \tag{4.111}$$

$$= \epsilon + \mathbb{E}_\sigma\left[\sup_{v'\in\mathcal{C}}\left(\frac{1}{n}\langle v', \sigma\rangle\right)\right] \tag{4.112}$$

$$\leq \epsilon + \sqrt{\frac{2\log|\mathcal{C}|}{n}} \qquad (\text{Proposition 4.6.1}) \tag{4.113}$$

$$= \epsilon + \sqrt{\frac{2\log N(\epsilon, \mathcal{Q}, \rho)}{n}} \tag{4.114}$$

$$= \epsilon + \sqrt{\frac{2\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \qquad (\text{Remark 4.24}) \tag{4.115}$$

Since the argument above holds for any $\epsilon > 0$, we can take the infimum over all ϵ to arrive at Equation (4.106). □

4.6.1 Chaining and Dudley's theorem

While Theorem 4.25 is useful, the bound in (4.108) is rarely tight as z might not be perfectly correlated with σ. It is possible to obtain a stronger theorem by constructing a "chained ϵ-covering scheme." Specifically, when we decompose $v = v' + z$, we can construct a finer-grained covering of the ball $B(v', \epsilon)$, and then we can decompose z into smaller components and so on (see Figure 4.5 for an illustration).

Using this method of chaining, we can obtain the following (stronger) result:

Theorem 4.26 (Dudley's Theorem)**.** *If $\mathcal{F}$ is a function class from $Z \mapsto \mathbb{R}$, then*

$$R_S(\mathcal{F}) \le 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon. \tag{4.116}$$

Note that unlike in Theorem 4.25, we do not require $f \in \mathcal{F}$ to be bounded.

It is not obvious how (4.116) improves upon the one-step discretization bound given by (4.106). At a high level, we can interpret this bound as removing the discretization error term by averaging over different scales of ϵ. But before we can explicitly prove this claim, we motivate our approach. In the proof of Theorem 4.25, we approximated v with $v' + z$ where v' is the closest point to v in the minimal ϵ-cover of $\mathcal{Q}$, and z is the vector between v' and v. In particular,

$$\frac{1}{n}\langle v, \sigma\rangle = \frac{1}{n}\langle v', \sigma\rangle + \frac{1}{n}\langle z, \sigma\rangle \tag{4.117}$$

Then, to obtain a bound, we take a sup of both sides, but apply the sup separately to each term on the right hand side. Namely, we show that:

$$\mathbb{E}\left[\sup_v \frac{1}{n}\langle v, \sigma\rangle\right] \le \mathbb{E}\left[\sup_{v' \in \mathcal{C}} \frac{1}{n}\langle v', \sigma\rangle\right] + \mathbb{E}\left[\sup_{z \in B_{v'}} \frac{1}{n}\langle z, \sigma\rangle\right] \tag{4.118}$$

This bound follows by observing that $\mathbb{E}[\sup(A+B)] \le \mathbb{E}[\sup A] + \mathbb{E}[\sup B]$ since the sup on the RHS is taken separately over both terms. The difficult term to tightly bound is the last one, $\frac{1}{n}\langle z, \sigma\rangle$. In the previous derivation, we naively upper bounded $\langle z, \sigma\rangle$ using Cauchy-Schwarz,

$$\frac{1}{n}\langle z, \sigma\rangle \le \frac{\|z\|_2 \cdot \|\sigma\|_2}{n}, \tag{4.119}$$

but this bound is only tight if there exists $z \in B_{v'}$ that is perfectly correlated with σ. We claim that such perfect correlation is unlikely. Recall that the output space is defined by possible outputs of $f \in \mathcal{F}$ given n inputs. Unless our function class is extremely expressive, the set of radius ϵ around v' contained in $\mathcal{Q}$ will only be a small subset of the ϵ-ball centered at v'; thus, $\sup_z \frac{1}{n}\langle z, \sigma\rangle \ll \frac{\|z\|_2 \cdot \|\sigma\|_2}{n}$.

To precisely set up our approach, we observe that $\mathbb{E}[\sup_{z \in B_{v'}} \frac{1}{n}\langle z, \sigma\rangle]$ is itself a Rademacher complexity: $R_S(B_{v'} \cap \mathcal{Q})$. To more tightly bound $\mathbb{E}\left[\sup_{z \in B_{v'}} \frac{1}{n}\langle z, \sigma\rangle\right]$, we then repeat the ϵ-covering argument again with a smaller choice of ϵ. Intuitively, this procedure amounts to "decomposing $\langle z, \sigma\rangle$ from (4.117) into another pair of terms corresponding to the new ϵ-cover and the discretization error." "Chaining" then repeats this decomposition countably many times. This procedure is illustrated visually by Figure 4.5, and we formalize this argument in the sequel.

Proof. Let $\epsilon_0 = \sup_{f \in \mathcal{F}} \max_i |f(z_i)|$, so that for all $v \in \mathcal{Q}$,

$$\epsilon_0 \ge \sqrt{\frac{1}{n}\sum_{i=1}^n f(z_i)^2} = \sqrt{\frac{1}{n}\|v\|_2^2}. \tag{4.120}$$

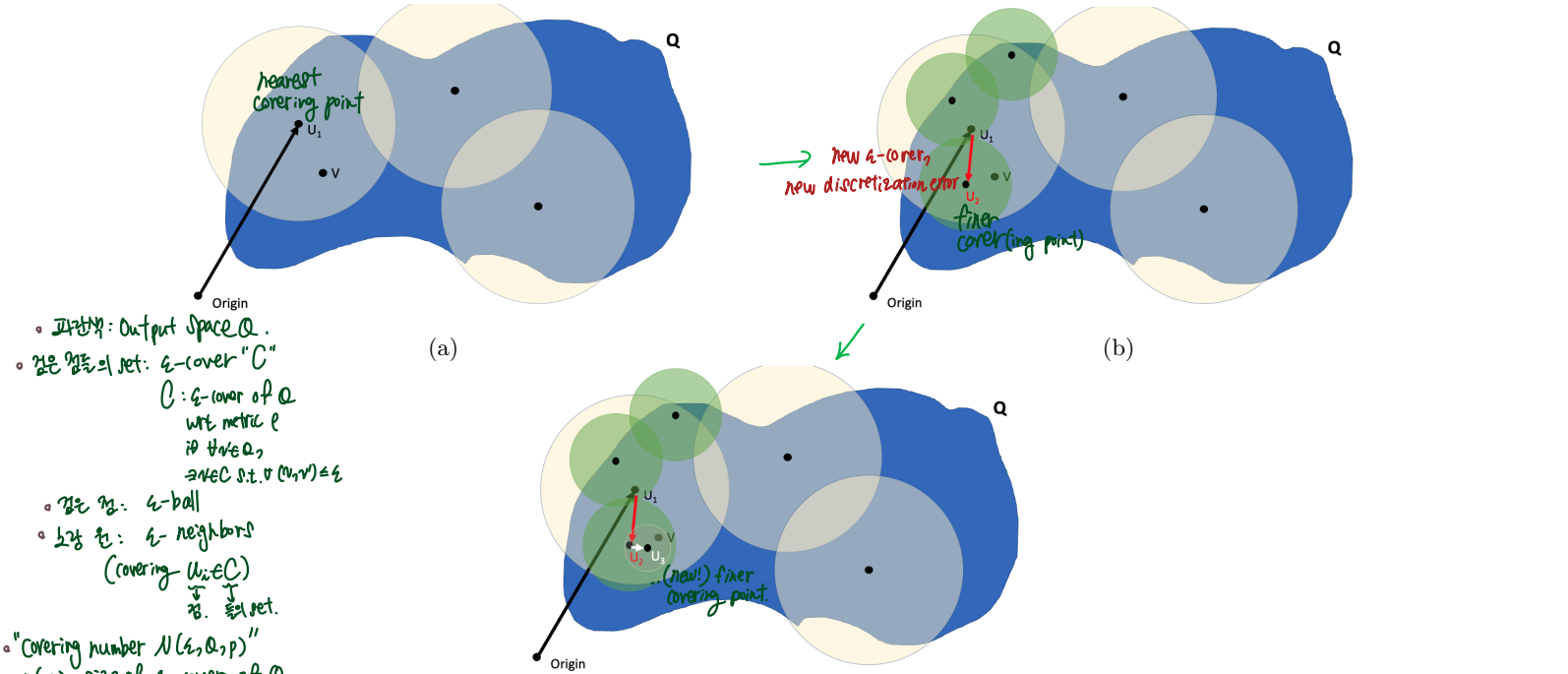

Figure 4.5: We depict how the chaining procedure approximates v using a sequence of progressively finer discretizations. Figure 4.5a illustrates how we first approximate v using the nearest covering point u_1, while Figures 4.5b and 4.5c describe how we refine this approximation using two finer covers, whose nearest points are denoted by u_2 and u_3, respectively.

Define $\epsilon_j = 2^{-j}\epsilon_0$ and let $\mathcal{C}_j$ be an ϵ_j-cover of $\mathcal{Q}$. Then, $\mathcal{C}_0$ is the coarsest cover of $\mathcal{Q}$, and as j increases, we obtain progressively more fine-grained covers $\mathcal{C}_j$. (We can intuitively think of these covers as nested, but this is not necessary for the proof to hold.) We next use this sequence of covers to define a telescoping series that equals v; the terms in this series can then be analyzed using the tools that we have developed in the prequel.

For $v \in \mathcal{Q}$, let u_i denote the nearest neighbor of v in $\mathcal{C}_i$. (Note that by definition $\rho(u, v_j) \leq \epsilon_j$.) Taking $u_0 = 0$, it follows from our definition of $\mathcal{C}_i$ that as $j \to \infty$, $\epsilon_j \to 0$ and $u_j \to v$. Leveraging these observations, we can express v using the following series:

$$v = u_1 + (u_2 - u_1) + (u_3 - u_2) + \cdots \tag{4.121}$$

$$= (u_1 - u_0) + (u_2 - u_1) + (u_3 - u_2) + \cdots \tag{4.122}$$

$$= \sum_{i=1}^{\infty} (u_i - u_{i-1}). \tag{4.123}$$

Substituting (4.123) in the Rademacher complexity we aim to bound, we obtain

$$\mathbb{E}\left[\sup_{v\in\mathcal{Q}}\frac{1}{n}\langle v,\sigma\rangle\right]=\mathbb{E}\left[\sup_{v\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{\infty}\langle u_i-u_{i-1},\sigma\rangle\right] \tag{4.124}$$

$$\leq\mathbb{E}\left[\sum_{i=1}^{\infty}\sup_{(u_i\in\mathcal{C}_i),(u_{i-1}\in\mathcal{C}_{i-1})}\frac{1}{n}\langle u_i-u_{i-1},\sigma\rangle\right] \tag{4.125}$$

$$=\sum_{i=1}^{\infty}\mathbb{E}\left[\sup_{(u_i\in\mathcal{C}_i),(u_{i-1}\in\mathcal{C}_{i-1})}\frac{1}{n}\langle u_i-u_{i-1},\sigma\rangle\right]. \tag{4.126}$$

Observe that

$$\mathbb{E}\left[\sup_{u_i\in\mathcal{C}_i,u_{i-1}\in\mathcal{C}_{i-1}}\frac{1}{n}\langle u_i-u_{i-1},\sigma\rangle\right] \tag{4.127}$$

is a Rademacher complexity defined over the *finite* space $\mathcal{C}_i\times\mathcal{C}_{i-1}$, so we can use Proposition 4.6.1 (Massart's lemma) to obtain a tractable upper bound. To do so, we must first compute an upper bound on $\frac{1}{\sqrt{n}}\|u_i-u_{i-1}\|_2$:

$$\frac{1}{\sqrt{n}}\|u_i-u_{i-1}\|_2=\frac{1}{\sqrt{n}}\|(u_i-v)-(u_{i-1}-v)\|_2 \tag{4.128}$$

$$\leq\frac{1}{\sqrt{n}}\left(\|u_i-v\|_2-\|u_{i-1}-v\|_2\right) \tag{4.129}$$

$$\leq\epsilon_i+\epsilon_{i-1} \tag{4.130}$$

$$=3\epsilon_i \qquad (\epsilon_{i-1}\triangleq 2\epsilon_i) \tag{4.131}$$

Now we apply Proposition 4.6.1 with $M=3\epsilon_i$ and $|\mathcal{Q}|=|\mathcal{C}_i\times\mathcal{C}_{i-1}|\left(\leq|\mathcal{C}_i|\cdot|\mathcal{C}_{i-1}|.\right)$

$$\mathbb{E}\left[\sup_{u_i\in\mathcal{C}_i,u_{i-1}\in\mathcal{C}_{i-1}}\frac{1}{n}\langle u_i-u_{i-1},\sigma\rangle\right]\leq\sqrt{\frac{2(3\epsilon_i)^2\log(|\mathcal{C}_i|\cdot|\mathcal{C}_{i-1}|)}{n}} \tag{4.132}$$

$$=\frac{3\epsilon_i}{\sqrt{n}}\sqrt{2(\log|\mathcal{C}_i|+\log|\mathcal{C}_{i-1}|)} \tag{4.133}$$

$$\leq\frac{6\epsilon_i}{\sqrt{n}}\sqrt{\log(|\mathcal{C}_i|)} \qquad (|\mathcal{C}_i|\geq|\mathcal{C}_{i-1}|) \tag{4.134}$$

Applying (4.134) to each term in (4.126) and substituting the covering number $N(\epsilon_i,\mathcal{F},L_2(P_n))$ for $|\mathcal{C}_i|$, we obtain the following upper bound on the Rademacher complexity:

$$\mathbb{E}\left[\sup_{v\in\mathcal{Q}}\frac{1}{n}\langle v,\sigma\rangle\right]\leq\sum_{i=1}^{\infty}\frac{6\epsilon_i}{\sqrt{n}}\sqrt{\log N(\epsilon_i,\mathcal{F},L_2(P_n))}. \tag{4.135}$$

Finally, we must relate (4.135) to the target upper bound of $12\int\frac{1}{\sqrt{n}}\sqrt{\log N(\epsilon,\mathcal{F},L_2(P_n))}d\epsilon$. Examining Figure 4.6, we can make two crucial observations. First, for sufficiently large ϵ, $\log N(\epsilon,\mathcal{F},L_2(P_n))=0$ since one point is sufficient to construct a cover. Second, we observe that

$$(\epsilon_i-\epsilon_{i+1})\sqrt{\log|\mathcal{C}_i|}\leq\int_{\epsilon_{i+1}}^{\epsilon_i}\sqrt{\log N(\epsilon,\mathcal{F},L_2(P_n))}d\epsilon \tag{4.136}$$

since the LHS of (4.136) is the area of the dotted rectangle illustrated in Figure 4.6 while the RHS is the area under the curve for that interval. Formally, this result is equivalent to observing that the right Riemann sum underestimates the integral for monotone decreasing functions f.

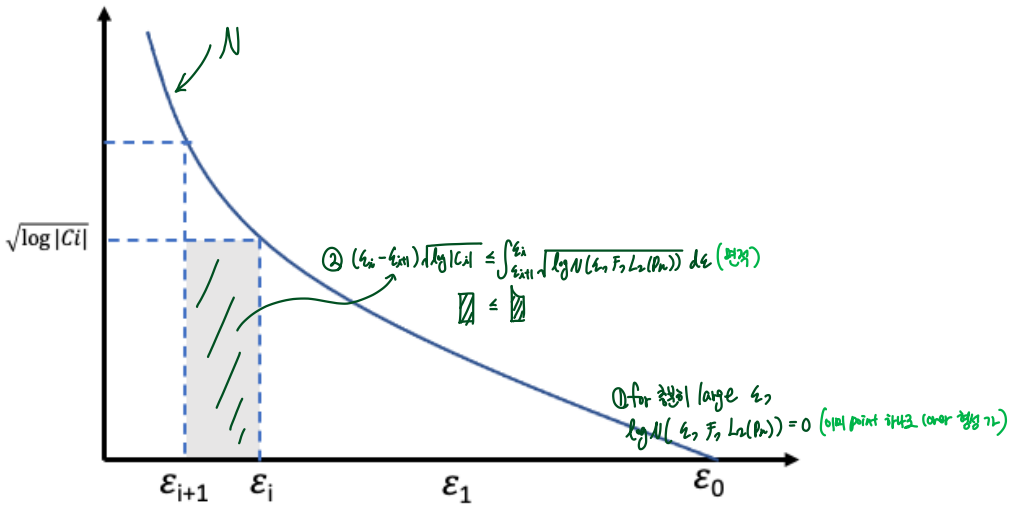

Figure 4.6: We observe that $\log N(\epsilon, \mathcal{F}, L_2(P_n))$ is monotone decreasing in ϵ. The area of the dotted rectangle formed by the vertical lines at ϵ_{i+1} and ϵ_i equals (up to a constant factor) the $i-$th term of the infinite sum derived in our proof of Dudley's theorem (4.135). The figure shows that the area of this rectangle is no larger than the integral of $\log N(\epsilon, \mathcal{F}, L_2(P_n))$ over this same interval.

Recognizing that $\epsilon_i - \epsilon_{i+1} = \frac{\epsilon_i}{2}$, we note that the LHS of (4.136) is equal (up to a constant factor) to the i-th term of (4.135). Thus,

$$\sum_{i=1}^{\infty} \frac{6\epsilon_i}{\sqrt{n}} \sqrt{\log N(\epsilon_i, \mathcal{F}, L_2(P_n))} = \frac{12}{\sqrt{n}} \sum_{i=1}^{\infty} (\epsilon_i - \epsilon_{i+1}) \sqrt{\log N(\epsilon_i, \mathcal{F}, L_2(P_n))} \tag{4.137}$$

$$\leq \frac{12}{\sqrt{n}} \int_{\epsilon_{i+1}}^{\epsilon_i} \sqrt{\log N(\epsilon_i, \mathcal{F}, L_2(P_n))} d\epsilon \tag{4.138}$$

$$= \frac{12}{\sqrt{n}} \int_{0}^{\epsilon_0} \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P_n))} d\epsilon. \tag{4.139}$$

To complete the proof, observe that $\log N(\epsilon, \mathcal{F}, L_2(P_n)) = 0$ for all $\epsilon > \epsilon_0$. This allows us to extend the upper limit of the integral given by (4.139) to ∞ and yields the desired result:

$$\mathbb{E}\left[\sup_{v \in \mathcal{Q}} \frac{1}{n} \langle v, \sigma \rangle\right] \leq \frac{12}{\sqrt{n}} \int_{0}^{\infty} \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P_n))} d\epsilon. \tag{4.140}$$

(from 4.116 to 4.140)

□

Remark 4.27. If $\mathcal{F}$ consists of functions bounded in $[-1, 1]$, then we have that for all $\epsilon > 1$, $N(\epsilon, \mathcal{F}, L_2(P_n)) = 1$. To see this, choose $\{f \equiv 0\}$, which is a complete cover for $\epsilon > 1$. Hence, the limits of integration in (4.116) can be truncated to $[0, 1]$:

$$R_S(\mathcal{F}) \leq 12 \int_{0}^{1} \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon, \tag{4.141}$$

since $\log N(\epsilon, \mathcal{F}, L_2(P_n)) = 0$ for $\epsilon > 1$.

Dudley's Theorem

4.6.2 Translating Covering Number Bounds to Rademacher Complexity

Of course, the bound in (4.116) is only useful if the integral on the RHS is finite. Here are some setups where this is the case (we continue to assume that the functions in $\mathcal{F}$ are bounded in $[-1,1]$):

1. If after ignoring multiplicative and additive constants,

$$N(\epsilon, \mathcal{F}, L_2(P_n)) \approx (1/\epsilon)^R, \tag{4.142}$$

then we have $\log N(\epsilon, \mathcal{F}, L_2(P_n)) \approx R \log(1/\epsilon)$. We can plug this into the RHS of (4.116) to get

$$\int_0^1 \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon = \int_0^1 \sqrt{\frac{R \log(1/\epsilon)}{n}} d\epsilon \approx \sqrt{\frac{R}{n}}. \tag{4.143}$$

2. If after ignoring multiplicative and additive constants, for some a,

$$N(\epsilon, \mathcal{F}, L_2(P_n)) \approx a^{R/\epsilon}, \tag{4.144}$$

then we have $\log N(\epsilon, \mathcal{F}, L_2(P_n)) \approx \frac{R}{\epsilon} \log a$. The bound in (4.116) becomes

$$\int_0^1 \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon \approx \int_0^1 \sqrt{\frac{R}{n\epsilon} \log a} d\epsilon \tag{4.145}$$

$$= \sqrt{\frac{R}{n} \log a} \int_0^1 \sqrt{\frac{1}{\epsilon}} d\epsilon \tag{4.146}$$

$$= \tilde{O}\left(\sqrt{\frac{R}{n}}\right). \tag{4.147}$$

3. If the covering number has the form $N(\epsilon, \mathcal{F}, L_2(P_n)) \approx a^{R/\epsilon^2}$, then $\log N(\epsilon, \mathcal{F}, L_2(P_n)) \approx \frac{R}{\epsilon^2} \log a$. In this case we have:

$$\int_0^1 \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon \approx \sqrt{\frac{R}{n} \log a} \int_0^1 \frac{1}{\epsilon} d\epsilon = \infty, \tag{4.148}$$

i.e. the bound in (4.116) is vacuous. This is because of the behavior of $\epsilon \mapsto 1/\epsilon^2$ near 0: the function goes to infinity too quickly for us to upper bound its integral. Fortunately, there is an "improved" version of Dudley's theorem that is applicable here:

Theorem 4.28 (Localized Dudley's Theorem)**.** *If $\mathcal{F}$ is a function class from $Z \mapsto \mathbb{R}$, then for any fixed cutoff $\alpha \geq 0$ we have the bound*

$$R_S(\mathcal{F}) \leq 4\alpha + 12 \int_{\alpha}^{\infty} \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon. \tag{4.149}$$

The proof of this theorem is similar to the proof of the original Dudley's theorem, except that the iterative covering procedure is stopped at the threshold $\epsilon = \alpha$, at the cost of the extra 4α term above.

Theorem 4.28 allows us to avoid the problematic region around $\epsilon = 0$ in the integral in (4.116). If we let $\alpha = 1/\mathsf{poly}(n)$, where $\mathsf{poly}(n)$ denotes some polynomial function of n, the bound in (4.149) becomes

$$R_S(\mathcal{F}) \le \frac{1}{\mathsf{poly}(n)} + \frac{\sqrt{R \log a}}{\sqrt{n}} \int_\alpha^1 \frac{1}{\epsilon} d\epsilon \tag{4.150}$$

$$= \frac{1}{\mathsf{poly}(n)} + \frac{\sqrt{R \log a}}{\sqrt{n}} \log(1/\alpha) \tag{4.151}$$

$$= \widetilde{O}\left(\sqrt{\frac{R}{n}}\right). \tag{4.152}$$

The last line follows by observing that $\log(1/\alpha) = \log \mathsf{poly}(n)$.

In summary, we have that $R_S(\mathcal{F}) \le \widetilde{O}\left(\sqrt{\frac{R}{n}}\right)$ for covering numbers of the form $R \log(1/\epsilon)$, $\frac{R}{\epsilon} \log a$, or $\frac{R}{\epsilon^2} \log a$ for some a. Note that if the dependence on ϵ is $1/\epsilon^c$ for $c > 2$, then even the improved Dudley's theorem does not help us. This is because the $\log(1/\alpha)$ term above becomes $\alpha^{1-c/2}$; then, for $\alpha = 1/\mathsf{poly}(n)$, the second term in Dudley's integral is no longer $\widetilde{O}\left(\sqrt{\frac{R}{n}}\right)$.

4.6.3 Lipschitz composition

Covering numbers also interact nicely with composition by Lipschitz functions. The following result is the analog of Talagrand's lemma for Rademacher complexity (Lemma 5.3), but its proof is much more elementary as given below. We will use this Lemma in Section 5.5 when bounding the covering number of deep nets.

Lemma 4.29. *Suppose ϕ is κ-Lipschitz, and $\rho = L_2(P_n)$. Then,*

$$\log N(\epsilon, \phi \circ \mathcal{F}, \rho) \le \log N(\epsilon/\kappa, \mathcal{F}, \rho) \tag{4.153}$$

Proof. Let $\mathcal{C}$ denote an ϵ/κ-cover for $\mathcal{F}$. Then $\phi \circ \mathcal{C}$ is an ϵ-cover of $\phi \circ \mathcal{F}$.

$$\rho(\phi \circ f', \phi \circ f) = \sqrt{\frac{1}{n} \sum (\phi(f'(z_i)) - \phi(f(z_i)))^2} \tag{4.154}$$

$$\le \sqrt{\frac{1}{n} \cdot \kappa^2 \sum (f'(z_i) - f(z_i))^2} \tag{4.155}$$

$$\le \kappa \cdot \frac{\epsilon}{\kappa} = \epsilon \tag{4.156}$$

□

4.7 VC dimension and its limitations

In this section, we briefly discuss a classical notion of complexity measure of function class, VC dimension. We will show that VC dimension is an upper bound on the Rademacher complexity. We will focus on classification and will be working within the framework of supervised learning stated in Chapter 1. The labels belong to the output space $\mathcal{Y} = \{-1, 1\}$, each classifier is a function $h : \mathcal{X} \to \mathbb{R}$ for all $h \in \mathcal{H}$, and the prediction is the sign of the output, i.e. $\hat{y} = \mathrm{sgn}(h(x))$. We will look at zero-one loss, i.e. $\ell_{0\text{-}1}((x, y), h) = \mathbb{1}(\mathrm{sgn}(h(x)) \neq y)$. Note that we can re-express the loss function as

$$\ell_{0\text{-}1}((x, y), h) = \frac{1 - \mathrm{sgn}(h(x))y}{2}. \tag{4.157}$$

The first approach is to reason directly about the Rademacher complexity of $\ell_{0\text{-}1}$ loss, i.e. considering the family of functions $\mathcal{F} = \{z = (x, y) \mapsto \ell_{0\text{-}1}((x, y), h) : h \in \mathcal{H}\}$. Define Q to be the set of all possible outputs

on our dataset: $Q = \{(\operatorname{sgn}(h(x^{(1)})), \ldots, \operatorname{sgn}(h(x^{(n)}))) \mid h \in \mathcal{H}\}$. Then, using our earlier remark about viewing the empirical Rademacher complexity as an inner product between $v \in Q$ and σ, we have

$$R_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1-\operatorname{sgn}(h(x^{(i)}))y_i}{2}\right] \tag{4.158}$$

$$= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{\operatorname{sgn}(h(x^{(i)}))}{2}\right] \tag{4.159}$$

$$= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{v\in Q} \frac{1}{n}\langle \sigma, v\rangle\right]. \tag{4.160}$$

Notice that the supremum is now over Q (instead of $\mathcal{F}$.) If n is sufficiently large, then it is typically the case that $|Q| > |\mathcal{F}|$. To see why this is the case, note that each function f corresponds to a single element in Q. However, as n increases, $|Q|$ increases as well. For any particular $v \in Q$, notice that $\langle v, \sigma\rangle$ is a sum of bounded random variables, so we can use Hoeffding's inequality to obtain

$$\Pr\left[\frac{1}{n}\langle \sigma, v\rangle \geq t\right] \leq \exp(-nt^2/2). \tag{4.161}$$

Taking the union bound over $v \in Q$, we see that

$$\Pr\left[\exists v \in Q \text{ such that } \frac{1}{n}\langle \sigma, v\rangle \geq t\right] \leq |Q|\exp(-nt^2/2). \tag{4.162}$$

Thus, with probability at least $1-\delta$, it is true that $\sup_{v\in Q} \frac{1}{n}\langle v, \sigma\rangle \leq \sqrt{\frac{2(\log|Q| + \log(2/\delta))}{n}}$. (Similarly, we can show that $\mathbb{E}\left[\sup_{v\in Q} \frac{1}{n}\langle v, \sigma\rangle\right] \leq O\left(\sqrt{\frac{\log|Q|+\log(2/\delta)}{n}}\right)$ holds.)

The key point to notice here is that the upper bound on $R_S(\mathcal{F})$ depends on $\log|Q|$. *VC dimension* is one way that we deal with bounding the size of Q. (We will not delve into the details of this approach (for those interested, see Section 3.11 of [Liang, 2016]).) VC dimension, however, has a number of limitations. For one, we will always end up with a bound that depends somehow on the dimension. For linear models, we obtain a bound $\log|Q| \lesssim d\log n$, corresponding to a bound on Rademacher complexity that looks like

$$R_S(\mathcal{F}) \leq \widetilde{O}\left(\sqrt{\frac{d}{n}}\right), \tag{4.163}$$

so we still have a $\sqrt{d}$ term. This will not be a good bound for high-dimensional models. For general models, we will arrive a bound of the form

$$R_S(\mathcal{F}) \leq \widetilde{O}\left(\sqrt{\frac{\#\text{ of parameters}}{n}}\right). \tag{4.164}$$

This upper bound only depends on the number of parameters in our model, and does not take into the account the scale and norm of the parameters. Additionally, this doesn't work with kernel methods since the explicit parameterization is possibly infinite-dimensional, and therefore this upper bound becomes useless.

These limitations motivate the use of "margin theory", which does take into account the norm of parameters and provides a theoretical basis for regularization techniques such as L_1 and L_2 regularization.