CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: Daniel Do

Lecture # 6
February 1st, 2021

6.1 Review and overview

In the previous lecture, we introduced the concepts of average Rademacher complexity and empirical Rademacher complexity. Here are their definitions again:

Definition 6.1 ((Average) 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

$$R_n(F) \triangleq \mathbb{E}_{z_1,\ldots,z_n \overset{\text{iid}}{\sim} P}\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n \overset{\text{iid}}{\sim} \{\pm 1\}}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]\right], \tag{6.1}$$

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

Definition 6.2 (Empirical Rademacher complexity)**.** Given a dataset $S = \{z_1, \ldots, z_n\}$, the *empirical Rademacher complexity* is defined as

$$R_S(\mathcal{F}) \triangleq \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]. \tag{6.2}$$

The following is the main theorem about Rademacher complexity that we derived last lecture.

Theorem 6.3. *Suppose for all $f \in \mathcal{F}$, $0 \le f(x) \le 1$. Then, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f]\right] \le 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{6.3}$$

Setting $\mathcal{F}$ to be a family of loss functions bounded by $[0,1]$ in the theorem gives the following corollary which we will make use of in this lecture:

Corollary 6.4. 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

$$L(h) - \hat{L}(h) \le 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{for all } h \in \mathcal{H}. \tag{6.4}$$

Remark 6.5. If we want to bound the generalization gap by the average Rademacher complexity instead, we can replace the RHS of (6.4) with $2R_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}$.

In this lecture, we will finish up with some intuition-forming remarks about Rademacher complexity and then proceed to use margin theory to utilize the Rademacher complexity to bound the generalization gap for the classification setting.

6.2 Some remarks about Rademacher complexity

6.2.1 Understanding the upper bound in Corollary 6.4

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 6.4 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 $h \in \mathcal{H}$. In previous lectures, 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.

6.2.2 Empirical Rademacher complexity as an inner product / viewed in the output space

Assume we have a fixed dataset $S = \{z_1, \ldots, z_n\}$. Since $z_1, \ldots, z_n$ is fixed, each function $f \in \mathcal{F}$ corresponds to a single output $(f(z_1), \ldots, f(z_n)) \in \mathbb{R}^n$. Hence, we can express the set of outputs for every function $f \in \mathcal{F}$ as

$$Q_{\mathcal{F}} = \left\{ \big(f(z_1), \ldots, f(z_n)\big) \mid f \in \mathcal{F} \right\}. \tag{6.5}$$

Now we can mathematically re-express the empirical Rademacher complexity as an inner product:

$$R_S(\mathcal{F}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] \tag{6.6}$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{v\in Q} \frac{1}{n}\langle \sigma, v\rangle\right], \tag{6.7}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$. (See Figure 6.1 for an illustration of this idea.) This perspective will be helpful later on when proving bounds on the empirical Rademacher complexity.

Another corollary of this is that the empirical Rademacher complexity doesn't depend on the exact parameterization of $\mathcal{F}$. For example, suppose we have two parameterizations $\mathcal{F} = \left\{f(x) = \sum \theta_i x_i \mid \theta \in \mathbb{R}^d\right\}$ and $\mathcal{F}' = \left\{f(x) = \sum \theta_i^3 \cdot w_i x_i \mid \theta \in \mathbb{R}^d, w \in \mathbb{R}^d\right\}$. 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 $Q_{\mathcal{F}}$.

6.2.3 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 6.1.)

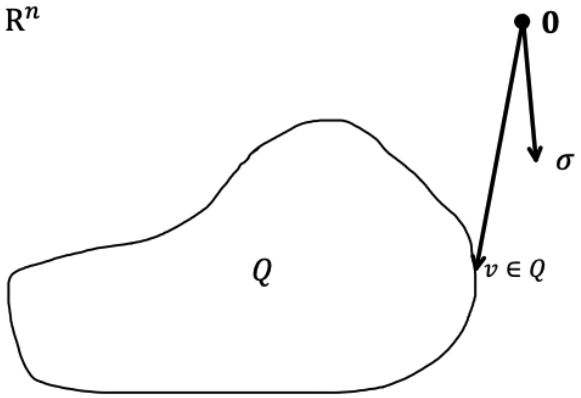

Figure 6.1: We can view empirical Rademacher complexity as the expectation of the maximum inner product between σ and $v \in Q$.

Proposition 6.2.1. 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.

$$R_S(\mathcal{F}') = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f' \in \mathcal{F}'} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] \tag{6.8}$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (f(z_i) + c_0)\right] \tag{6.9}$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i c_0 + \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] \tag{6.10}$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right] = R_S(\mathcal{F}), \tag{6.11}$$

where (6.11) 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. □

6.3 Motivation: VC dimension and its limitations

Now we will instantiate Rademacher complexity for specific cases. We will focus on classification and will be working within the framework of supervised learning stated in Lecture 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. $\hat{y} = \text{sgn}(h(x))$. We will look at zero-one loss, i.e. $\ell_{0\text{-}1}((x, y), h) = \mathbb{1}(\text{sgn}(h(x)) \neq y)$. Note that we can re-express the loss function as

$$\ell_{0\text{-}1}((x, y), h) = \frac{1 - \text{sgn}(h(x))y}{2}. \tag{6.12}$$

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

$$R_S(\mathcal{F}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \operatorname{sgn}(h(x^{(i)}))y_i}{2}\right] \tag{6.13}$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{\operatorname{sgn}(h(x^{(i)}))}{2}\right] \tag{6.14}$$

$$= \frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{v\in Q} \frac{1}{n}\langle \sigma, v\rangle\right]. \tag{6.15}$$

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

$$\Pr\left[\frac{1}{n}(\sigma, v) \geq t\right] \leq \exp(-nt^2/2). \tag{6.16}$$

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

$$\Pr\left[\exists v \in Q \text{ such that } \frac{1}{n}\langle \sigma, v\rangle \geq t\right] \leq |Q| \exp(-nt^2/2). \tag{6.17}$$

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 [Lia16]). 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

$$R_S(\mathcal{F}) \leq \widetilde{O}\left(\sqrt{\frac{d}{n}}\right), \tag{6.18}$$

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

$$R_S(\mathcal{F}) \leq \widetilde{O}\left(\sqrt{\frac{\#\text{ of parameters}}{n}}\right). \tag{6.19}$$

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 motivation 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.

6.4 Margin Theory

6.4.1 Intuition

Assume that we are in the same setting as in the previous section. A fundamental problem we face in this setting is that we do not have a continuous loss: everything is discrete in the output space. We need to find a way to reason about the scale of the output. An example of this is logistic regression: the logistic regression model outputs a probability, and while we compare it to the outcome (0 or 1), how close it is to the true output gives us a measure of how confident we are in the prediction.

Figure 6.2 gives similar intuition for linear classifiers. Intuitively, the black line is a "better" decision boundary than the red line because the minimum distance from any point to the black boundary is greater than the minimum distance from any point to the red line. In the next section, we will formalize this intuition by proving that the larger this margin is, the smaller the bound on generalization gap is.

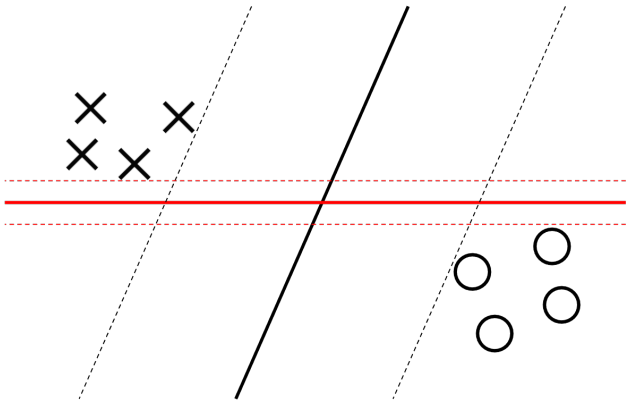

Figure 6.2: The red and black lines are two decision boundaries. The X's are positive examples and the O's are negative examples. The black line has a larger margin than the red line, and is intuitively a better classifier.

6.4.2 Formalizing margin theory

First, assume that the dataset $\mathcal{D} = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$ is *completely separable*. In other words, there exists some $h_\theta \in \mathcal{H}$ such that $y^{(i)} = \text{sgn}(h_\theta(x^{(i)}))$ holds for all $(x^{(i)}, y^{(i)}) \in \mathcal{D}$. This is not a necessary condition for our final bound but will make the derivation cleaner.

Definition 6.6 ((Unnormalized) Margin)**.** Fix the hypothesis h_θ. The *(unnormalized) margin* for example (x, y) is defined as $\text{margin}(x) = yh_\theta(x)$. Margin is only defined on examples where $\text{sgn}(h_\theta(x)) = y$. (Note that $\text{margin}(x) \geq 0$ because of our assumption of complete separability.)

Definition 6.7 (Minimum margin)**.** Given a dataset $\mathcal{D} = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$, the *minimum margin* over the dataset is defined as $\gamma_{\min} \triangleq \min_{i \in \{1,\ldots,|\mathcal{D}|\}} y^{(i)} h_\theta(x^{(i)})$.

Our final bound will have the form (generalization gap)$\leq f(\text{margin}, \text{parameter norm})$. This is very generic since there are many different bounds we could derive based on what margin we use. For this current setting we are using $\gamma_{\min}$, which is the minimum margin, but in other settings could use γ_{average}, which is the average margin of each point in the dataset.

We will begin by introducing the idea of a *surrogate loss*, a loss function which approximates zero-one loss but takes the scale of the margin into account. The *margin loss* (also known as *ramp loss*) is defined as

$$\ell_\gamma(t) = \begin{cases} 0 & t \geq \gamma \\ 1 & t \leq 0 \\ 1 - t/\gamma & 0 \leq t \leq \gamma \end{cases} \tag{6.20}$$

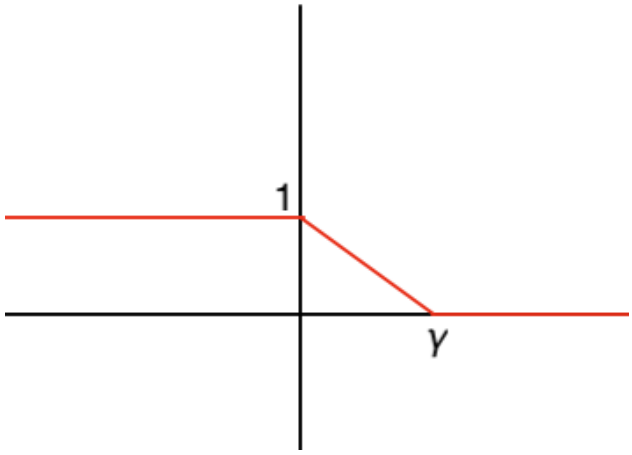

Figure 6.3: Plotted margin loss.

It is plotted in Figure 6.3. For convenience, define $\ell_\gamma((x, y), h) \triangleq \ell_\gamma(yh(x))$. We can view ℓ_γ as a continuous version of $\ell_{0\text{-}1}$ while being more sensitive to the scale of the margin on $[0, \gamma]$. Notice that $\ell_{0\text{-}1}$ is always less than or equal to the ℓ_γ when $\gamma \geq 0$, i.e.

$$\ell_{0\text{-}1}((x, y), h) \leq \ell_\gamma((x, y), h) \tag{6.21}$$

holds for all $(x, y) \sim P$. Taking the expectation over (x, y) on both sides of this inequality, we see that

$$L(h) = \mathbb{E}_{(x,y)\sim P}\left[\ell_{0\text{-}1}((x, y), h)\right] \leq \mathbb{E}_{(x,y)\sim P}\left[\ell_\gamma((x, y), h)\right]. \tag{6.22}$$

Therefore, the population loss is bounded by the expectation of the margin loss, and so it is sufficient to bound the expectation of the margin loss in order to bound the population loss.

Define the population and empirical version of the margin loss:

$$L_\gamma(h) = \mathbb{E}_{(x,y)\sim P}\left[\ell_\gamma((x,y),h)\right], \quad \hat{L}_\gamma(h) = \sum_{i=1}^{n}\left[\ell_\gamma((x^{(i)},y^{(i)}),h)\right]. \tag{6.23}$$

By Corollary 6.4, we see that with probability at least $1-\delta$ that

$$L_\gamma(h) - \hat{L}_\gamma(h) \le 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{6.24}$$

where $\mathcal{F} = \{(x,y) \mapsto \ell_\gamma((x,y),h) \mid h \in \mathcal{H}\}$. Note that if we set $\gamma \le \gamma_{\min}$, then $\hat{L}_\gamma(h) = 0$. This follows because by definition of $\gamma_{\min}$, $y^{(i)}h(x^{(i)}) \ge \gamma_{\min}$ for any $(x^{(i)},y^{(i)}) \in \mathcal{D}$. As a result, $\ell_\gamma((x^{(i)},y^{(i)}),h) = \ell_\gamma(y^{(i)}h(x^{(i)})) = 0$ holds. Therefore, it suffices to bound $R_S(\mathcal{F})$.

We will now use *Talagrand's lemma* to bound $R_S(\mathcal{F})$ in terms of $R_S(\mathcal{H})$ to remove any dependence on the loss function from the upper bound.

Lemma 6.8. *(Talagrand's lemma) Let $\phi : \mathbb{R} \to \mathbb{R}$ be a κ-Lipschitz function. Then*

$$R_S(\phi \circ \mathcal{H}) \le \kappa R_S(\mathcal{H}), \tag{6.25}$$

where $\phi \circ \mathcal{H} = \{z \mapsto \phi(h(z)) \mid h \in \mathcal{H}\}$.

We can use Talagrand's lemma directly with $\phi(t) = \ell_\gamma(t)$, which is $\frac{1}{\gamma}$-Lipschitz. We can express $\mathcal{F}$ as $\mathcal{F} = \ell_\gamma \circ \mathcal{H}'$ where $\mathcal{H}' = \{(x,y) \to yh(x) \mid h \in \mathcal{H}\}$. Applying Talagrand's lemma, we see that

$$\begin{aligned}
R_S(\mathcal{F}) &\le \frac{1}{\gamma}R_S(\mathcal{H}') && (6.26)\\
&= \frac{1}{\gamma}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i y^{(i)}h(x^{(i)})\right] && (6.27)\\
&= \frac{1}{\gamma}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(x^{(i)})\right] && (6.28)\\
&= \frac{1}{\gamma}R_S(\mathcal{H}). && (6.29)
\end{aligned}$$

Putting this all together, we have shown that for $\gamma \le \gamma_{\min}$,

$$\begin{aligned}
L_{0\text{-}1}(h) \le L_\gamma(h) &\le 0 + O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) + \widetilde{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right) && (6.30)\\
&= O\left(\frac{R_S(\mathcal{H})}{\min_i y^{(i)}h(x^{(i)})}\right) + \widetilde{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right). && (6.31)
\end{aligned}$$

In the next lecture, we will talk about how we can bound $R_S(\mathcal{H})$ by the norms for linear models, two-layer neural networks and multi-layer neural networks.

Remark 6.9. Note there is a subtlety here. If we think of the dataset as random, it follows that $\gamma_{\min}$ is a random variable. Consequently, the γ we choose to define the hypothesis class is random, which is not a valid choice when thinking about Rademacher complexity! Technically we cannot apply Talagrand's lemma with a random κ (which we took to be $1/\gamma$). Also, when we used concentration inequalities, we implicitly assume that the $\ell_\gamma((x^{(i)}, y^{(i)}), h)$ are independent of each other. That is not the case if γ is dependent on the data.

How can we address this? The idea is to do another union bound over γ. Choose a family $\Gamma = \left\{2^k : k \in [-B, B]\right\}$ for some B. For every fixed $\gamma \in \Gamma$, we prove the theorem that

$$L_{0\text{-}1}(h) \leq \widehat{L}_\gamma(h) + O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{6.32}$$

We can then take a union bound over all $\gamma \in \Gamma$. Next, choose the largest $\gamma \in \Gamma$ such that $\gamma \leq \gamma_{\min}$. For this value of γ we have $\widehat{L}_\gamma(h) = 0$ and $O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) = O\left(\frac{R_S(\mathcal{H})}{\gamma_{\min}}\right)$.

Bibliography

[Lia16] Percy Liang, *Cs229t/stat231: Statistical learning theory (winter 2016)*, April 2016.