CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: Spencer M. Richards and Thomas Lew

Lecture # 7
Feb. 3rd, 2021

7.1 Review and overview

In the previous lecture, we discussed margin theory, margin loss, and how to use it to bound the generalization gap for binary classifiers. We found that for training data of the form $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n \subset \mathbb{R}^d \times \{-1, 1\}$, a hypothesis class $\mathcal{H}$ and 0-1 loss, we could derive a bound of the form

$$\text{generalization loss} \leq \frac{2R_S(\mathcal{H})}{\gamma_{\min}} + \text{low-order term}, \tag{7.1}$$

where $\gamma_{\min}$ is the minimum margin achievable on S over those hypotheses in $\mathcal{H}$ that separate the data, and $R_S(\mathcal{H})$ is the empirical Rademacher complexity of $\mathcal{H}$. Such bounds state that simpler models will generalize better beyond the training data, particularly for data that is strongly separable.

In this lecture, we study bounds on the Rademacher complexity of two hypothesis classes: linear models and two-layer neural networks.

7.2 Rademacher complexity of linear models

7.2.1 Linear models with weights bounded in ℓ_2 norm

We begin with the Rademacher complexity of linear models using weights with bounded ℓ_2 norm.

Theorem 7.1. *Let $\mathcal{H} = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \leq B\}$ for some constant $B > 0$. Moreover, assume $\mathbb{E}_{x\sim P}\left[\|x\|_2^2\right] \leq C^2$, where P is some distribution and $C > 0$ is a constant. Then*

$$R_S(\mathcal{H}) \leq \frac{B}{n}\sqrt{\sum_{i=1}^n \left\|x^{(i)}\right\|_2^2}, \tag{7.2}$$

and

$$R_n(\mathcal{H}) \leq \frac{BC}{\sqrt{n}}. \tag{7.3}$$

Generally speaking, there are two methods with which we can bound the Rademacher complexity of a model. The first method, which we used in the last lecture, consists of discretizing the space of possible outputs from our hypothesis class, then using a union bound or covering number argument to bound the Rademacher complexity of the model. While this method is powerful and generally applicable, it yields bounds that depend on the logarithm of the cardinality of this discretized output space, which in turn depends on the number of data points n. In the proof below, we will instead use a more elegant, albeit limited technique which does not rely on discretization of the output space.

Proof. We start with the proof of (7.2). By definition,

$$R_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{\|w\|_2 \le B} \frac{1}{n} \sum_{i=1}^n \sigma_i \left\langle w, x^{(i)} \right\rangle \right] \tag{7.4}$$

$$= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\|w\|_2 \le B} \left\langle w, \sum_{i=1}^n \sigma_i x^{(i)} \right\rangle \right] \tag{7.5}$$

$$= \frac{B}{n} \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x^{(i)} \right\|_2 \right] \qquad (\sup_{\|w\|_2 \le B} \langle w, v \rangle = B \|v\|_2) \tag{7.6}$$

$$\le \frac{B}{n} \sqrt{\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x^{(i)} \right\|_2^2 \right]} \qquad (\text{Jensen's ineq. for } \alpha \mapsto \alpha^2) \tag{7.7}$$

$$= \frac{B}{n} \sqrt{\mathbb{E}_\sigma \left[\sum_{i=1}^n \left(\sigma_i^2 \left\| x^{(i)} \right\|_2^2 + \left\langle \sigma_i x^{(i)}, \sum_{j \ne i}^n \sigma_j x^{(j)} \right\rangle \right) \right]} \tag{7.8}$$

$$= \frac{B}{n} \sqrt{\sum_{i=1}^n \left\| x^{(i)} \right\|_2^2}. \qquad (\sigma_i \text{ indep. and } \mathbb{E}[\sigma_i] = 0) \tag{7.9}$$

This completes the proof of (7.2) for the empirical Rademacher complexity. The bound on the average Rademacher complexity in (7.3) follows from taking the expectation of both sides to get

$$R_n(\mathcal{H}) = \mathbb{E}\left[R_S(\mathcal{H})\right] = \frac{B}{n} \mathbb{E}\left[\sqrt{\sum_{i=1}^n \left\| x^{(i)} \right\|_2^2} \right] \le \frac{B}{n} \sqrt{\sum_{i=1}^n \mathbb{E}\left[\left\| x^{(i)} \right\|_2^2 \right]} \le \frac{BC}{\sqrt{n}}, \tag{7.10}$$

where the first inequality is another application of Jensen's inequality, and the second follows from the assumption $\mathbb{E}_{x \sim P}\left[\|x\|_2^2\right] \le C^2$.

□

We observe that both the empirical and average Rademacher complexities scale with the upper ℓ_2-norm bound $\|w\|_2 \le B$ on the parameters w, which motivates regularizing the model. However, smaller weights in the model may reduce the margin $\gamma_{\min}$, which in turn hurts generalization according to (7.1).

Remark 7.2. Note that if we scale the data by some multiplicative factor, the bound on empirical Rademacher complexity $R_S(\mathcal{H})$ will scale accordingly. However, at the same time we expect the margin to scale by the same multiplicative factor, so the bound on the generalization gap in (7.1) does not change. This lines up with our intuition that the bound should not depend on the scaling of the data.

7.2.2 Linear models with weights bounded in ℓ_1 norm

Now, we consider linear models again, except we restrict the ℓ_1-norm of the parameters and assume an ℓ_∞-norm bound on the data.

Theorem 7.3. *Let $\mathcal{H} = \left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_1 \le B\right\}$ for some constant $B > 0$. Moreover, assume $\left\|x^{(i)}\right\|_\infty \le C$ for some constant $C > 0$ and all points in $S = \{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$. Then*

$$R_S(\mathcal{H}) \le BC\sqrt{\frac{2\log(2d)}{n}}. \tag{7.11}$$

To prove the theorem, we will need Massart's lemma, which provides a bound for the Rademacher complexity of a finite hypothesis class.

Lemma 7.4 (Massart's lemma)**.** *Suppose $\mathcal{Q} \subset \mathbb{R}^n$ is finite and contained in the ℓ_2-norm ball of radius $M\sqrt{n}$ for some constant $M > 0$, i.e.,*

$$\mathcal{Q} \subset \{v \in \mathbb{R}^n \mid \|v\|_2 \le M\sqrt{n}\}. \tag{7.12}$$

Then, for Rademacher variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbb{R}^n$,

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathcal{Q}} \frac{1}{n}\langle\sigma, v\rangle\right] \le M\sqrt{\frac{2\log|\mathcal{Q}|}{n}}. \tag{7.13}$$

As a corollary, if $\mathcal{F}$ is a set of real-valued functions satisfying

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z^{(i)})^2 \le M^2, \tag{7.14}$$

over some data $S = \{z^{(i)}\}_{i=1}^n$, then

$$R_S(\mathcal{F}) \le M\sqrt{\frac{2\log|\mathcal{F}|}{n}}, \quad \textit{and} \quad R_n(\mathcal{F}) \le M\sqrt{\frac{2\log|\mathcal{F}|}{n}}. \tag{7.15}$$

We will not prove Massart's lemma in detail. The intuition is to use concentration inequalities to bound $\frac{1}{n}\langle\sigma, v\rangle$ for fixed v, then to use a union bound over the elements $v \in \mathcal{Q}$.

We will now prove Theorem 7.3:

Proof of Theorem 7.3. By definition,

$$\begin{aligned} R_S(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_{\|w\|_1\le B} \frac{1}{n}\sum_{i=1}^n \sigma_i\left\langle w, x^{(i)}\right\rangle\right] && (7.16)\\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|w\|_1\le B}\left\langle w, \sum_{i=1}^n \sigma_i x^{(i)}\right\rangle\right] && (7.17)\\ &= \frac{B}{n}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i x^{(i)}\right\|_\infty\right], && (7.18) \end{aligned}$$

where the last equality is because $\sup_{\|w\|_1\le B}\langle w, v\rangle = B\|v\|_\infty$, i.e., the ℓ_∞-norm is the dual of the ℓ_1-norm, which is a consequence of Hölder's inequality. However, the ℓ_∞-norm is difficult to simplify further. Instead, we use the fact that $\sup_{\|w\|_1\le 1}\langle w, v\rangle$ for any $v \in \mathbb{R}^d$ is always attained at

one of the vertices $\mathcal{W} = \bigcup_{i=1}^{d} \{-e_i, e_i\}$, where $e_i \in \mathbb{R}^d$ is the i-th coordinate unit vector. Defining the restricted hypothesis class $\bar{\mathcal{H}} = \{x \mapsto \langle w, x \rangle \mid w \in \mathcal{W}\} \subset \mathcal{H}$, this yields

$$R_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\|w\|_1 \leq B} \left\langle w, \sum_{i=1}^{n} \sigma_i x^{(i)} \right\rangle \right] \tag{7.19}$$

$$= \frac{B}{n}\mathbb{E}_\sigma \left[\max_{w \in \mathcal{W}} \left\langle w, \sum_{i=1}^{n} \sigma_i x^{(i)} \right\rangle \right] \tag{7.20}$$

$$= BR_S(\bar{\mathcal{H}}). \tag{7.21}$$

Since $\bar{\mathcal{H}} \subset \mathcal{H}$, necessarily $R_S(\bar{\mathcal{H}}) \leq R_S(\mathcal{H})$. In particular, the model class $\bar{\mathcal{H}}$ is bounded and finite with cardinality $|\bar{\mathcal{H}}| = 2d$. This suggests using Massart's lemma to complete the proof. To do so, we need to confirm that $\bar{\mathcal{H}}$ is bounded with respect to the ℓ_2-metric. Indeed, since the inner product of $x^{(i)}$ with a coordinate vector e_j just selects the j-th coordinate of $x^{(i)}$, for any $w \in \mathcal{W}$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \left\langle w, x^{(i)} \right\rangle^2 \leq \frac{1}{n}\sum_{i=1}^{n} \left\| x^{(i)} \right\|_\infty^2 \leq \frac{1}{n}\sum_{i=1}^{n} C^2 = C^2, \tag{7.22}$$

where the last inequality uses the assumption $\|x_i\|_\infty \leq C$. So $\bar{\mathcal{H}}$ is bounded in the ℓ_2-metric and finite, thus by Massart's Lemma we have

$$R_S(\mathcal{H}) = BR_S(\bar{\mathcal{H}}) \leq BC\sqrt{\frac{2\log|\bar{\mathcal{H}}|}{n}} = BC\sqrt{\frac{2\log(2d)}{n}}, \tag{7.23}$$

which completes the proof. □

7.2.3 Comparing the bounds for different $\mathcal{H}$

First, we note that for this hypothesis class of linear models, it is possible to obtain an upper bound proportional to $\sqrt{d/n}$ using the VC dimension, which grows quickly with the data dimension d. Our bound is better since it does not have as strong of a dependence on d, and accounts for the norms of our model parameters and the data.

In the two subsections above, we considered two different hypothesis classes of linear models, each restricting different norms. In both cases, the bound on the average Rademacher complexity depended on the product of the norm bound on the parameters w and the norm bound on each data point x. To determine which choice of hypothesis class is better, consider the bounds

$$\|w\|_2 \|x\|_2 \quad \text{vs.} \quad \|w\|_1 \|x\|_\infty$$

and see how they compare in different settings. We consider 3 settings here:

- Suppose w and x are random variables with w_i and x_i close to the set of values $\{-1, 1\}$. Then we have

  $$\sqrt{d} \cdot \sqrt{d} \quad \text{vs.} \quad d \cdot 1.$$

  In this case, there is no difference in using either linear hypothesis class.

- If we additionally suppose w is sparse with at most k non-zero entries, then we have

$$\sqrt{k} \cdot \sqrt{d} \quad \text{vs.} \quad k \cdot 1.$$

So for $d \gg k$, we have $\sqrt{kd} \gg k$ and thus ℓ_1-norm regularization leads to a better complexity bound when w is suspected to be sparse. Indeed, $\sqrt{d}\,\|x\|_\infty \approx \|x\|_2$ when the entries of x are somewhat uniformly distributed, and so in the sparse case we have

$$\|w\|_2 \|x\|_2 \geq \sqrt{d}\,\|w\|_2 \|x\|_\infty \geq \|w\|_1 \|x\|_\infty . \tag{7.24}$$

- On the other hand, if w is dense in the sense that $\|w\|_2 \approx \sqrt{d}\,\|w\|_1$ (i.e., if all entries in w are close to each other in magnitude), then

$$\|w\|_2 \|x\|_2 \leq \frac{1}{\sqrt{d}} \|w\|_1 \cdot \sqrt{d}\,\|x\|_\infty \leq \|w\|_1 \|x\|_\infty . \tag{7.25}$$

In this case, it makes sense to regularize the ℓ_2-norm instead.

In practice, other multiplicative factors enter the generalization bound, so regularizing both the ℓ_1- and ℓ_2-norms of the model parameters w is preferable.

Continuing with this rough style of analysis, for the hypothesis class with restricted ℓ_2-norm, we can write the bound on the generalization gap in (7.1) as

$$\text{generalization loss} \lesssim \frac{\|w\|_2 \|x\|_2}{\sqrt{n}\gamma_{\min}} + \text{low-order term}. \tag{7.26}$$

The presence of $\|w\|_2 / \gamma_{\min}$ motivates both the minimum norm and the maximum margin formulations of the Support Vector Machine (SVM) problem as good methods to improve generalization performance of binary classifiers.

7.3 Rademacher complexity of two-layer neural networks

We now compute a bound for the Rademacher complexity of two-layer neural networks. Throughout this section, we use the following notation:

- $\theta = (w, U)$ are the parameters of the model with $w \in \mathbb{R}^m$ and $U \in \mathbb{R}^{m \times d}$, where m denotes the number of hidden units. We use $u_i \in \mathbb{R}^d$ to denote the i-th row of U (written as a column vector).
- $\phi(z) = \max(z, 0)$ is the ReLU activation function applied element-wise.
- $f_\theta(x) = \langle w, \phi(Ux) \rangle = w^\top \phi(Ux)$ is the model.
- $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is the training set, with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$.

We start with a somewhat weak bound which introduces the technical tools we need to derive tighter bounds in subsequent lectures.

Theorem 7.5. *For some constants $B_w > 0$ and $B_u > 0$, let*

$$\mathcal{H} = \{f_\theta \mid \|w\|_2 \le B_w,\ \|u_i\|_2 \le B_u,\ \forall i \in \{1, 2, \dots, m\}\}, \tag{7.27}$$

and suppose $\mathbb{E}\left[\|x\|_2^2\right] \le C^2$. Then

$$R_n(\mathcal{H}) \le 2B_w B_u C\sqrt{\frac{m}{n}}. \tag{7.28}$$

This bound is not ideal as it depends on the number of neurons m. Empirically, it has been found that the generalization error does *not* increase monotonically with m. As more neurons are added to the model, thereby giving it more expressive power, studies have shown that generalization is improved [BHMM19]. This contradicts the bound above, which states that more neurons leads to worse generalization. Nevertheless, we now derive this bound.

Proof. By definition,

$$\begin{aligned}
R_S(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_\theta \frac{1}{n}\sum_{i=1}^n \sigma_i \left\langle w, \phi(Ux^{(i)})\right\rangle\right] && (7.29)\\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{U:\|u_j\|_2\le B_u}\ \sup_{\|w\|_2\le B_w} \left\langle w, \sum_{i=1}^n \sigma_i\phi(Ux^{(i)})\right\rangle\right] && (7.30)\\
&= \frac{B_w}{n}\mathbb{E}_\sigma\left[\sup_{U:\|u_j\|_2\le B_u} \left\|\sum_{i=1}^n \sigma_i\phi(Ux^{(i)})\right\|_2\right] && (\sup_{\|w\|_2\le B}\langle w, v\rangle = B\|v\|_2) \quad (7.31)\\
&\le \frac{B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{U:\|u_j\|_2\le B_u} \left\|\sum_{i=1}^n \sigma_i\phi(Ux^{(i)})\right\|_\infty\right] && (\|v\|_2 \le m\|v\|_\infty) \quad (7.32)\\
&= \frac{B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{U:\|u_j\|_2\le B_u}\ \max_{1\le j\le m} \left|\sum_{i=1}^n \sigma_i\phi(u_j^\top x^{(i)})\right|\right] && (7.33)\\
&= \frac{B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u} \left|\sum_{i=1}^n \sigma_i\phi(u^\top x^{(i)})\right|\right] && (7.34)\\
&\le \frac{2B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u} \sum_{i=1}^n \sigma_i\phi(u^\top x^{(i)})\right] && (\text{to be explained later}) \quad (7.35)\\
&\le \frac{2B_w\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\le B_u} \sum_{i=1}^n \sigma_i u^\top x^{(i)}\right], && (7.36)
\end{aligned}$$

where the last inequality follows by applying the contraction lemma (Talagrand's lemma) and observing that the ReLU function is 1-Lipschitz. (Observe that the expectation in (7.35) is the Rademacher complexity for $\{x \mapsto \phi(u^\top x) \mid \|u\|_2 \le B_u\}$: this is the family that we are applying the contraction lemma to.)

We now observe that the expectation in (7.36) is the Rademacher complexity of the family of linear models $\{x \mapsto \langle u, x\rangle \mid \|u\|_2 \le B_u\}$. Thus, applying Theorem 7.3 yields

$$R_S(\mathcal{H}) \le \frac{2B_w\sqrt{m}}{n} B_u \sqrt{\sum_{i=1}^n \left\|x^{(i)}\right\|_2^2}. \tag{7.37}$$

Taking the expectation of both sides and using similar steps to those in the proof of Theorem 7.3 gives us

$$R_n(\mathcal{H}) = \mathbb{E}\left[R_S(\mathcal{H})\right] \tag{7.38}$$

$$\leq \frac{2B_w B_u \sqrt{m}}{n} \mathbb{E}\left[\sqrt{\sum_{i=1}^{n} \left\|x^{(i)}\right\|_2^2}\right] \tag{7.39}$$

$$\leq \frac{2B_w B_u \sqrt{m}}{n} C\sqrt{n} \tag{7.40}$$

$$= 2B_w B_u C\sqrt{\frac{m}{n}}, \tag{7.41}$$

which completes the proof.

□

Next, we look at a finer bound that results from defining a new complexity measure:

Theorem 7.6. *Let* $C(\theta) = \sum_{j=1}^{m} |w_j| \, \|u_j\|_2$*, and for some constant* $B_C > 0$ *consider the hypothesis class*

$$\mathcal{H} = \{f_\theta \mid C(\theta) \leq B_C\}. \tag{7.42}$$

If $\left\|x^{(i)}\right\|_2 \leq C$ *for all* $i \in \{1, \ldots, n\}$*, then*

$$R_S(\mathcal{H}) \leq \frac{2B_C C}{\sqrt{n}}. \tag{7.43}$$

We defer the proof of this theorem to the next lecture. We conclude the lecture we some remarks:

Remark 7.7. We can think of $C(\theta)$ as a new complexity measure. One nice thing about hypothesis classes defined by $C(\theta)$ is that it is invariant to transformations of the form $(w, U) \mapsto (kw, U/k)$ for some constant $k > 0$: for the ReLU activation function ϕ, we have

$$f_{(w,U)}(x) = w^\top \max(Ux, 0) = kw^\top \max\left(\frac{1}{k}Ux, 0\right) = f_{(kw,U/k)}(x). \tag{7.44}$$

This is not the case for the (implicit) complexity measure in Theorem 7.5: these transformations would change the values of B_w and B_u.

Remark 7.8. Compared to Theorem 7.5, this bound does not explicitly depend on the number of neurons m. Thus, it is possible to use more neurons and still maintain a tight bound if the value of the new complexity measure $C(\theta)$ is reasonable. In contrast, the bound of Theorem 7.5 only looks at the total number of neurons. With the result above, it is possible to regularize $C(\theta)$ and obtain a tighter bound for any number m of neurons. This would lead to an accurate model with better generalization guarantees for a high number neurons.

For example, consider solving the constrained problem

$$\rho_m = \min_\theta C(\theta) \quad \text{such that} \quad f_\theta \text{ fits the data } \{(x^{(i)}, y^{(i)})\}_{i=1}^n. \tag{7.45}$$

In this case, ρ_m monotonically decreases as the number of neurons m increases. Indeed, models with more parameters necessarily include models with a lower number of parameters and thus those of lower complexity. As a result, it is possible to obtain lower complexity models by increasing the number of parameters m.

Bibliography

[BHMM19] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias-variance trade-off*, Proceedings of the National Academy of Sciences (PNAS) **116** (2019), no. 32, 15849–15854.