CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: David Lin and Jinhui Wang

Lecture # 8
Feb. 8th, 2021

8.1 Review and overview

In the previous lecture, we derived explicit Rademacher complexity bounds for linear models with weights bounded in ℓ_1 and ℓ_2 norms. To recap, we proved:

Theorem 8.1. *Let $\mathcal{H} = \left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\right\}$ for some constant $B > 0$. Moreover, assume $\mathbb{E}_{x\sim P}\left[\|x\|_2^2\right] \le C^2$, where P is some distribution and $C > 0$ is a constant. Then*

$$R_S(\mathcal{H}) \le \frac{B}{n}\sqrt{\sum_{i=1}^{n}\left\|x^{(i)}\right\|_2^2}, \quad \text{and} \quad R_n(\mathcal{H}) \le \frac{BC}{\sqrt{n}}. \tag{8.1}$$

Theorem 8.2. *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 $S = \left\{x^{(i)}\right\}_{i=1}^n \subset \mathbb{R}^d$. Then*

$$R_S(\mathcal{H}) \le BC\sqrt{\frac{2\log(2d)}{n}}. \tag{8.2}$$

We also derived a crude bound for a two-layer neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}$ defined by $f_\theta(x) = w^\top\phi(Ux)$. Here, $U = \begin{pmatrix} u_1^\top \\ \vdots \\ u_m^\top \end{pmatrix} \in \mathbb{R}^{m\times d}$ and $w \in \mathbb{R}^m$ correspond to the weight matrices in the first and second layers of the neural network respectively while $\phi(z) = \max(z, 0)$ is the ReLU activation function taken elementwise. (We can think of the network as being parameterized by the pair $\theta = (w, U)$.) For this set-up, we showed:

Theorem 8.3. *For some constants $B_w > 0$ and $B_u > 0$, let*

$$\mathcal{H} = \left\{f_\theta \mid \|w\|_2 \le B_w, \|u_i\|_2 \le B_u, \forall i \in \{1, 2, \ldots, m\}\right\}, \tag{8.3}$$

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

$$R_n(\mathcal{H}) \le 2B_wB_uC\sqrt{\frac{m}{n}}. \tag{8.4}$$

This upper bound is undesirable since it grows with the number of neurons m, contradicting empirical observations of the generalization error decreasing with m. Today, we will complete the proof of a refined upper bound that is non-increasing with m by leveraging a finer definition of complexity $C(\theta)$. Subsequently, we discuss implications of this new theorem before transitioning to a new method of bounding Rademacher complexities by covering the output space.

8.2 Refined bound for two-layer neural networks

A recurring theme in subsequent proofs will be the functional invariance of two-layer neural networks under a class of rescaling transformations. The key ingredient will be the *positive homogeneity* of the ReLU function, i.e.

$$\alpha\phi(x) = \phi(\alpha x) \qquad \forall \alpha > 0. \tag{8.5}$$

This implies that for any $\lambda_i > 0$ $(i = 1, \ldots, m)$, the transformation $\theta = \{(w_i, u_i)\}_{1 \le i \le m} \mapsto \theta' = \{(\lambda_i w_i, u_i/\lambda_i)\}_{1 \le i \le m}$ has no net effect on the neural network's functionality (i.e. $f_\theta = f_{\theta'}$) since

$$w_i \cdot \phi\left(u_i^\top x^{(i)}\right) = (\lambda_i w_i) \cdot \phi\left(\left(\frac{u_i}{\lambda_i}\right)^\top x^{(i)}\right). \tag{8.6}$$

In light of this, we devise a new complexity measure $C(\theta)$ that is also invariant under such transformations and use it to prove a better bound for the Rademacher complexity. This positive homogeneity property is absent in the (implicit) complexity measure used in the hypothesis class (8.3) of Theorem 8.3.

Theorem 8.4. Let $C(\theta) = \sum_{j=1}^m |w_j| \, \|u_j\|_2$, *and for some constant* $B_C > 0$ *consider the hypothesis class*

$$\mathcal{H} = \{f_\theta \mid C(\theta) \le B_C\}. \tag{8.7}$$

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

$$R_S(\mathcal{H}) \le \frac{2B_C C}{\sqrt{n}}. \tag{8.8}$$

Proof. Due to the positive homogeneity of the ReLU function ϕ, it will be useful to define the ℓ_2-normalized weight vector $\bar{u}_j := u_j/\|u_j\|_2$ so that $\phi\left(u_j^T x\right) = \|u_j\|_2 \cdot \phi(\bar{u}_j^T x)$. The empirical Rademacher complexity satisfies

$$R_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_\theta \sum_{i=1}^n \sigma_i f_\theta\left(x^{(i)}\right)\right] \tag{8.9}$$

$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_\theta \sum_{i=1}^n \sigma_i \left[\sum_{j=1}^m w_j \phi\left(u_j^T x^{(i)}\right)\right]\right] \qquad \text{(by dfn of } f_\theta\text{)} \tag{8.10}$$

$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_\theta \sum_{i=1}^n \sigma_i \left[\sum_{j=1}^m w_j \|u_j\|_2 \phi\left(\bar{u}_j^T x^{(i)}\right)\right]\right] \qquad \text{(by positive homogeneity of } \phi\text{)} \tag{8.11}$$

$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_\theta \sum_{j=1}^m w_j \|u_j\|_2 \left[\sum_{i=1}^n \sigma_i \phi\left(\bar{u}_j^T x^{(i)}\right)\right]\right] \tag{8.12}$$

$$\le \frac{1}{n}\mathbb{E}_\sigma\left[\sup_\theta \sum_{j=1}^m |w_j| \|u_j\|_2 \max_{k \in [n]} \left|\sum_{i=1}^n \sigma_i \phi\left(\bar{u}_k^T x^{(i)}\right)\right|\right] \quad \left(\because \sum_j \alpha_j \beta_j \le \sum_j |\alpha_j| \max_k |\beta_k|\right) \tag{8.13}$$

$$\leq \frac{B_C}{n}\mathbb{E}_\sigma\left[\sup_{\theta=(w,U)}\max_{k\in[n]}\left|\sum_{i=1}^n \sigma_i\phi\left(\bar{u}_k^T x^{(i)}\right)\right|\right] \qquad (\because C(\theta)\leq B_C) \tag{8.14}$$

$$= \frac{B_C}{n}\mathbb{E}_\sigma\left[\sup_{\bar{u}:\|\bar{u}\|_2=1}\left|\sum_{i=1}^n \sigma_i\phi\left(\bar{u}^T x^{(i)}\right)\right|\right] \tag{8.15}$$

$$\leq \frac{B_C}{n}\mathbb{E}_\sigma\left[\sup_{\bar{u}:\|\bar{u}\|_2\leq 1}\left|\sum_{i=1}^n \sigma_i\phi\left(\bar{u}^T x^{(i)}\right)\right|\right] \tag{8.16}$$

$$\leq \frac{2B_C}{n}\mathbb{E}_\sigma\left[\sup_{\bar{u}:\|\bar{u}\|_2\leq 1}\sum_{i=1}^n \sigma_i\phi\left(\bar{u}^T x^{(i)}\right)\right] \qquad \text{(see Lemma 8.6)} \tag{8.17}$$

$$= 2B_C R_S(\mathcal{H}'), \tag{8.18}$$

where $\mathcal{H}' = \left\{x \mapsto \phi(\bar{u}^T x) : \bar{u}\in\mathbb{R}^d, \|\bar{u}\|_2\leq 1\right\}$. By Talagrand's lemma, since ϕ is 1-Lipschitz, $R_S(\mathcal{H}')\leq R_S(\mathcal{H}'')$ where $\mathcal{H}'' = \left\{x\mapsto \bar{u}^T x : \bar{u}\in\mathbb{R}^d, \|\bar{u}\|_2\leq 1\right\}$ is a linear hypothesis space. Using $R_S(\mathcal{H}'')\leq \frac{C}{\sqrt{n}}$ by Theorem 8.1 then concludes the proof.

□

Remark 8.5. This refined upper bound is much stronger than Theorem 8.3 since it does not increase with the number of neurons m. Notably, we can analyse the $m\to\infty$ limit and still hope to obtain a non-vacuous bound. Intuitively, this is because the network has a fixed "budget" B_C that it is allowed to distribute among an arbitrary number of neurons.

We complete the proof by deriving the Lemma 8.6 used in the second last inequality. Notably, the lemma's assumption holds in the current context, since

$$\sup_\theta\langle\sigma, f_\theta(x)\rangle = \sup_{\bar{u}:\|\bar{u}\|_2\leq 1}\sum_{i=1}^n \sigma_i\phi\left(\bar{u}^T x^{(i)}\right)\geq 0. \tag{8.19}$$

since one can take $\bar{u}=0$ for any $\sigma=(\sigma_1,\ldots,\sigma_n)$.

Lemma 8.6. *Let* $\sigma=(\sigma_1,...,\sigma_n)$ *and* $f_\theta(x) = \left(f_\theta\left(x^{(1)}\right),...,f_\theta\left(x^{(n)}\right)\right)$. *Suppose that for any* $\sigma\in\{\pm1\}^n$, $\sup_\theta\langle\sigma,f_\theta(x)\rangle\geq 0$. *Then,*

$$\mathbb{E}_\sigma\left[\sup_\theta|\langle\sigma,f_\theta(x)\rangle|\right]\leq 2\mathbb{E}_\sigma\left[\sup_\theta\langle\sigma,f_\theta(x)\rangle\right]. \tag{8.20}$$

Proof. Letting ϕ be the ReLU function, the lemma's assumption implies that $\sup_\theta\phi\left(\langle\sigma,f_\theta(x)\rangle\right) = \sup_\theta\langle\sigma,f_\theta(x)\rangle$ for any $\sigma\in\{\pm1\}^n$. Observing that $|z| = \phi(z)+\phi(-z)$,

$$\sup_\theta|\langle\sigma,f_\theta(x)\rangle| = \sup_\theta\left[\phi\left(\langle\sigma,f_\theta(x)\rangle\right)+\phi\left(\langle-\sigma,f_\theta(x)\rangle\right)\right] \tag{8.21}$$

$$\leq \sup_\theta\phi\left(\langle\sigma,f_\theta(x)\rangle\right)+\sup_\theta\phi\left(\langle-\sigma,f_\theta(x)\rangle\right) \tag{8.22}$$

$$= \sup_\theta\langle\sigma,f_\theta(x)\rangle+\sup_\theta\langle-\sigma,f_\theta(x)\rangle. \tag{8.23}$$

Taking the expectation over σ (and noting that $\sigma\overset{d}{=}-\sigma$), we get the desired conclusion. □

8.3 Consequences of refined bound

In this section, we discuss practical implications of the refined neural network bound.

8.3.1 Connection to ℓ_2 regularization

Recall that margin theory yields

$$\text{for all } \theta, \quad L_{0\text{-}1}(\theta) \leq \frac{2R_S(\mathcal{H})}{\gamma_{\min}} + \widetilde{O}\left(\sqrt{\frac{\log(2/\delta)}{n}}\right), \tag{8.24}$$

with probability at least $1-\delta$. Thus, Theorem 8.4 motivates us to minimize $\frac{R_S(\mathcal{H})}{\gamma_{\min}}$ by regularizing $C(\theta)$. Concretely, this can be formulated as the optimization problem

$$\begin{aligned} \text{minimize} \quad & C(\theta) = \sum_{j=1}^{m} |w_j| \cdot \|u_j\|_2 \\ \text{subject to} \quad & \gamma_{\min}(\theta) \geq 1, \end{aligned} \tag{I}$$

or equivalently,

$$\begin{aligned} \text{maximize} \quad & \gamma_{\min}(\theta) \\ \text{subject to} \quad & C(\theta) \leq 1. \end{aligned} \tag{II}$$

At first glance, the above seems orthogonal to techniques used in practice. However, it turns out that the optimal neural network from (I) is functionally equivalent to that of the new problem:

$$\begin{aligned} \text{minimize} \quad & C_{\ell_2}(\theta) = \frac{1}{2}\sum_{j=1}^{m} |w_j|^2 + \frac{1}{2}\sum_{j=1}^{m} \|u_j\|_2^2 \\ \text{subject to} \quad & \gamma_{\min}(\theta) \geq 1. \end{aligned} \tag{I*}$$

This is a simple consequence of the positive homogeneity of ϕ. For any scaling factor $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}_+^m$, the rescaled neural network $\theta_\lambda := \{(\lambda_i w_i, u_i/\lambda_i)\}$ has the same functionality as the original neural network $\theta = \{w_i, u_i\}$ (i.e. it achieves the same $\gamma_{\min}$). Thus,

$$\min_\theta C_{\ell_2}(\theta) = \min_\theta \min_\lambda \left(\frac{1}{2}\sum_{j=1}^{m} \lambda_j^2 |w_j|^2 + \frac{1}{2}\sum_{j=1}^{m} \lambda_j^{-2} \|u_j\|_2^2\right) \tag{8.25}$$

$$= \min_\theta \sum_{j=1}^{m} |w_j| \cdot \|u_j\|_2 \tag{8.26}$$

$$= \min_\theta C(\theta) \tag{8.27}$$

where we have used the equality case of the AM-GM inequality, attainable by $\lambda_j^* = \sqrt{\frac{\|u_j\|_2}{|w_j|}}$, in the second step. This equality case also shows that $\theta^* = \{(w_i, u_i)\}$ is the optimal solution of (I) if and only if $\hat{\theta}^* = \theta_{\lambda^*}$ is the optimal solution of (I*)—proving that $\hat{\theta}^*$ and θ^* are functionally equivalent since they only differ by a positive scale factor.

This connects our $C(\theta)$ regularization to ℓ_2-norm penalties that are more prevalent in practice. In retrospect, we see this equivalence is essentially due to the positive homogeneity of the neural network which "homogenizes" any inhomogeneous objective such as C_{ℓ_2}. Hence, we can just deal with $C(\theta)$ which is transparently homogeneous.

8.3.2 Stable generalization bound in m

Next, we show that the generalization bound given by Theorem 8.4 does not deteriorate with the network width (number of neurons) m, which is consistent with experimental results. To this end, the perspective of (II) enables us to isolate all dependencies of m in $\gamma_{\min}$. Letting $\widehat{\theta}_m$ denote the minimizer of program (II) with width m and defining optimal value $\gamma_m^* = \gamma_{\min}\left(\widehat{\theta}_m\right)$, we can rewrite the margin bound (8.24) as

$$L(\widehat{\theta}_m) \leq \frac{4C}{\sqrt{n}} \cdot \frac{1}{\gamma_m^*} + (\text{lower-order terms}), \tag{8.28}$$

where all dependencies on m are now contained in γ_m^*. Hence, to show that this bound does not worsen as m grows, we just have to show that γ_m^* is non-decreasing in m. This is intuitively the case since a neural network of width $m+1$ contains one of width m under the same complexity constraints. The following theorem formalizes this hunch:

Theorem 8.7. *Let γ_m^* be the minimum margin obtained by solving* (II) *with a two-layer neural network of width m. Then $\gamma_m^* \leq \gamma_{m+j}^*$ for all positive integers j.*

Proof. Suppose $\theta = \{(w_i, u_i)\}_{1 \leq i \leq m}$ is a two-layer neural network of width m satisfying $C(\theta) \leq 1$. Then we may construct a neural network $\widetilde{\theta} = \{(\tilde{w}_i, \tilde{u}_i)\}_{1 \leq i \leq m+1}$ of width $m+1$ by simply taking

$$(\widetilde{w}_i, \widetilde{u}_i) = \begin{cases} (w_i, u_i) & i \leq m, \\ (0, 0) & \text{otherwise} \end{cases} \tag{8.29}$$

$\widetilde{\theta}$ is functionally equivalent to θ and $C(\widetilde{\theta}) = C(\theta) \leq 1$. This means maximizing $\gamma_{\min}$ over $\{C(\widetilde{\theta}) : \widetilde{\theta} \text{ of width } m+1\}$ should give no lower of a value than the maximum of $\gamma_{\min}$ over $\{C(\theta) : \theta \text{ of width } m\}$. □

8.3.3 Equivalence to an ℓ_1-SVM in $m \to \infty$ limit

Since γ_m^* is non-decreasing in m, quantity

$$\gamma_\infty^* = \lim_{m \to \infty} \gamma_m^* \tag{8.30}$$

is well-defined. The next interesting fact is that in this $m \to \infty$ limit, γ_∞^* of the two-layer neural network is equivalent to the minimum margin of an ℓ_1-SVM. As a brief digression, we recap the formulation of ℓ_p-SVMs and discuss the importance of ℓ_1-SVMs in particular.

Since a collection of data points with binary class labels may not be a priori separable, a *kernel model* first transforms an input x to $\Phi(x)$ where $\Phi : \mathbb{R}^d \to \mathcal{G}$ is known as the *feature map*. The model then seeks a separating hyperplane in this new (extremely high-dimensional) feature space $\mathcal{G}$, parameterized by a vector μ pointing from the origin to the hyperplane. The prediction of the

model on an input x is then a decision score that quantifies $\Phi(x)$'s displacement with respect to the hyperplane:

$$g_{\mu,\Phi}(x) := \langle \mu, \Phi(x) \rangle . \tag{8.31}$$

Motivated by margin theory, it is desirable to seek the maximum-margin hyperplane under a constraint on μ to guarantee the generalizability of the model. In particular, a kernel model with an ℓ_p-constraint seeks to solve the following program:

$$\begin{aligned} \text{maximize} \qquad & \gamma_{min} := \min_{i \in [n]} y^{(i)} \langle \mu, \Phi(x^{(i)}) \rangle \\ \text{subject to} \qquad & \|\mu\|_p \leq 1. \end{aligned} \tag{8.32}$$

Observe that both the prediction and optimization of the feature model only rely on inner products in $\mathcal{G}$. The ingenuity of the SVM is to choose maps Φ such that $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ can be directly computed in terms of x and x' in the original space $\mathbb{R}^d$, thereby circumventing the need to perform expensive inner products in the large space $\mathcal{G}$. Remarkably, this "kernel trick" enables us to even operate in an implicit, infinite-dimensional $\mathcal{G}$.

The case of $p = 1$ is particularly useful in practice as ℓ_1-regularization generally produces sparse feature weights (the constrained parameter space is a polyhedron and the optimum tends to lie at one of its vertices). Hence, ℓ_1-regularization is an important feature selection method when one expects only a few dimensions of $\mathcal{G}$ to be significant. Unfortunately, the ℓ_1-SVM is not kernelizable due to the kernel trick relying on ℓ_2-geometry, and is hence infeasible to implement. However, our next theorem shows that a two-layer neural network can approximate a particular ℓ_1-SVM in the $m \to \infty$ limit (and in fact, for finite m). For the sake of simplicity, we sacrifice rigor in defining the space $\mathcal{G}$ and convey the main ideas.

Theorem 8.8. *Define the feature map $\phi_{\text{relu}} : \mathbb{R}^d \to \mathcal{G}$ such that x is mapped to $\phi(u^\top x)$ for all vectors u on the $d-1$-dimensional sphere $\mathcal{S}^{d-1}$. Informally,*

$$\phi_{\text{relu}}(x) := \begin{bmatrix} \vdots \\ \phi(u^\top x) \\ \vdots \end{bmatrix}_{u \in S^{d-1}}$$

is an "infinite-dimensional vector" that contains an entry $\phi(u^T x)$ for each vector $u \in \mathcal{S}^{d-1}$, and we let $\phi_{\text{relu}}(x)[u]$ denote the "u"-th entry of this vector. Noting that $\mathcal{G}$ is the space of functions which can be indexed by $u \in S^{d-1}$, the inner product structure on $\mathcal{G}$ is defined by $\langle f, g \rangle = \int_{S^{d-1}} f[u] g[u] du$.

Under this set-up, we have

$$\gamma^*_\infty = \gamma^*_{\ell_1}, \tag{8.33}$$

where $\gamma^*_{\ell_1}$ is the minimum margin of the optimized ℓ_1-SVM with $\Phi = \phi_{\text{relu}}$.

Proof. We will only prove the $\gamma^*_\infty \leq \gamma^*_{\ell_1}$ direction. (The $\gamma^*_\infty \geq \gamma^*_{\ell_1}$ direction is essentially the same but requires slightly more work.)

Suppose γ^*_∞ is obtained by network weights $(w_1, w_2, \cdots), (u_1, u_2, \cdots)$ where $w_i \in \mathbb{R}, u_i \in \mathbb{R}^d$ (there is a slight subtlety here to be rectified later). Define renormalized versions of $\{w_i\}$ and $\{u_i\}$:

$$\widetilde{w}_i := w_i \cdot \|u_i\|_2, \qquad \overline{u}_i := \frac{u_i}{\|u_i\|_2}. \tag{8.34}$$

Note that $\{(\widetilde{w}_i, \overline{u}_i)\}$ has the same functionality (and also the same complexity measure $C(\theta)$, margin, etc.) as that of $\{(w_i, u_i)\}$, but now $\overline{u}_i$ has unit ℓ_2-norm (i.e. $\overline{u}_i \in \mathcal{S}^{d-1}$). Thus, $\phi(\overline{u}_i^\top x)$ can be treated as a feature in $\mathcal{G}$ and we can construct an equivalent ℓ_1-SVM (denoted by μ) such that $\widetilde{w}_i$ is the coefficient of μ associated with that feature. Since $\widetilde{w}_i$ must only be "turned on' at $\overline{u}_i$, we have

$$\mu[u] = \sum_{i \in \mathcal{S}^{d-1}} \tilde{w}_i \delta(u - \overline{u}_i), \tag{8.35}$$

where $\delta(u)$ is the Dirac-delta function. Given this μ, we can check that the SVM's prediction is

$$\begin{aligned} g_{\mu,\phi_{\text{relu}}}(x) &= \int_{S^{d-1}} \mu[u]\phi_{\text{relu}}(x)[u]du && (8.36)\\ &= \int_{S^{d-1}} \sum_{i \in \mathcal{S}^{d-1}} \tilde{w}_i \delta(u - \overline{u}_i)\phi\left(\overline{u}^\top x\right) du && (8.37)\\ &= \sum_{i \in \mathcal{S}^{d-1}} \tilde{w}_i \phi\left(\overline{u}_i^\top x\right), && (8.38) \end{aligned}$$

which is identical to the output $f_{\{(\widetilde{w}_i, \overline{u}_i)\}}(x)$ of the neural network. Furthermore,

$$\|\mu\|_1 = \sum_{i=1}^{\infty} |\widetilde{w}_i| = \sum_{i=1}^{\infty} |w_i| \cdot \|u_i\|_2 \leq 1, \tag{8.39}$$

where the last equality holds because $\{(\widetilde{w}_i, \overline{u}_i)\}$ satisfies the constraints of (II). This shows that our constructed μ satisfies the ℓ_1-SVM constraint. Thus, $\gamma^*_\infty \leq \gamma^*_{\ell_1}$ since the functional behavior of the optimal neural network is contained in the search range of the SVM.

□

Remark 8.9. How well does a finite dimensional neural network approximate the infinite-dimensional ℓ_1 network? Proposition B.11 of [WLLM20] shows that you only need $n+1$ neurons. Another way to say this is that $\{\gamma_m\}$ stabilizes once $m = n+1$:

$$\gamma^*_1 \leq \gamma^*_2 \leq \cdots \leq \gamma^*_{n+1} = \gamma^*_\infty. \tag{8.40}$$

The main idea of the proof is that if we have a neural net θ with $(n+2)$ neurons, then we can always pick a simplification θ' with $(n+1)$ neurons such that θ, θ' agree on all n datapoints.

As an aside, this result also resolves the issue in our partial proof. A priori, γ^*_∞ may not necessarily be attained by a set of weights $\{(\widetilde{w}_i, \overline{u}_i)\}$ but the above shows that it is achievable with just $n+1$ non-zero indices.

8.4 Covering number approach for Rademacher complexity

Our previous Rademacher complexity bounds hinged 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 $z_1, ..., z_n$,

$$R_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right], \tag{8.41}$$

we only care about each function's $f \in \mathcal{F}$ behavior on $\{z_1, ..., z_n\}$. Hence, we can forget the rest of the input space and characterize $f \in \mathcal{F}$ by its outputs $(f(z_1), \ldots, f(z_n))$. Thus, there is a paradigm shift from the space of all functions $\mathcal{F}$ to the *output space*

$$\mathcal{Q} \triangleq \left\{ \big(f(z_1), \ldots, f(z_n)\big)^\top : f \in \mathcal{F} \right\} \subseteq \mathbb{R}^n, \tag{8.42}$$

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

$$R_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{v \in \mathcal{Q}} \frac{1}{n} \langle \sigma, v \rangle \right]. \tag{8.43}$$

Now, for finite $|\mathcal{Q}|$, we immediately obtain the following bound by Massart's lemma:

$$R_S(\mathcal{F}) \leq \sqrt{\frac{2 \log |\mathcal{Q}|}{n}}. \tag{8.44}$$

When $|\mathcal{Q}|$ is infinite, we can use the same discretization trick that we used to prove the generalization bound for an infinite-hypothesis space. Instead of trying to cover the parameter space, we try to cover the output space. To this end, we firstly recall a few definitions concerning ϵ-covers.

Definition 8.10. $\mathcal{C}$ 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 8.11. The *covering number* is defined as the minimum size of an ϵ-cover, or explicitly:

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

The standard metric we will use is $\rho(v, v') = \frac{1}{\sqrt{n}} \|v - v'\|_2$, with the leading coefficient inserted for convenience.

Remark 8.12. While we want to consider ϵ-covers over $\mathcal{Q}$, the notation in the literature refers to them as ϵ-covers of the function class $\mathcal{F}$ using the metric $\rho = L_2(p_n)$, i.e.

$$\rho(f, f') = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f(z_i) - f'(z_i))^2} \tag{8.45}$$

If we take the corresponding $v, v' \in \mathcal{Q}$, this is precisely $\rho(v, v') = \frac{1}{\sqrt{n}} \|v - v'\|_2$.

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

Theorem 8.13. *Let $\mathcal{F}$ be a family of functions $Z \mapsto [-1, 1]$. Then*

$$R_S(\mathcal{F}) \leq \inf_{\epsilon > 0} \left(\epsilon + \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \right). \tag{8.46}$$

The ϵ term can be thought of as the discretization error, while the latter term is the term from Massart's lemma.

Proof. Fix any $\epsilon > 0$. Let $\mathcal{C}$ be an ϵ-cover $\mathcal{C}$ of $\mathcal{Q}$. Massart's lemma immediately gives the bound

$$R_S(\mathcal{C}) \leq \sqrt{\frac{2 \log |\mathcal{C}|}{n}}. \tag{8.47}$$

For every point $v \in \mathcal{Q}$, we can express it as $v = v' + z$, where $v' \in \mathcal{C}$ and z is small (specifically, $\frac{1}{\sqrt{n}}\|z\|_2 \leq \epsilon$). This gives

$$\begin{aligned}
\frac{1}{n}\langle v, \sigma \rangle &= \frac{1}{n}\langle v', \sigma \rangle + \frac{1}{n}\langle z, \sigma \rangle && (8.48)\\
&\leq \frac{1}{n}\langle v', \sigma \rangle + \frac{1}{n}\|z\|_2 \|\sigma\|_2 \qquad \text{(Cauchy-Schwarz)} && (8.49)\\
&\leq \frac{1}{n}\langle v', \sigma \rangle + \epsilon. \qquad \text{(since } \|z\|_2 \leq \sqrt{n}\epsilon \text{ and } \|\sigma\|_2 \leq \sqrt{n}\text{)} && (8.50)
\end{aligned}$$

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

$$\begin{aligned}
R_S(\mathcal{F}) &= \mathbb{E}_\sigma \left[\sup_{v \in \mathcal{Q}} \frac{1}{n}\langle v, \sigma \rangle \right] && (8.51)\\
&\leq \mathbb{E}_\sigma \left[\sup_{v' \in \mathcal{C}} \left(\frac{1}{n}\langle v', \sigma \rangle + \epsilon \right) \right] && (8.52)\\
&\leq \sqrt{\frac{2 \log |\mathcal{C}|}{n}} + \epsilon. \qquad \text{(Massart's lemma)} && (8.53)
\end{aligned}$$

Choosing $\mathcal{C}$ to be a minimal ϵ-cover allows us to set $|\mathcal{C}| = N(\epsilon, \mathcal{F}, L_2(p_n))$. Since the argument above holds for any $\epsilon > 0$, we can take the infimum over all ϵ to arrive at Equation (8.46), completing the proof.

□

While this theorem is useful, the bound in Equation (8.49) 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 8.1 for an illustration).

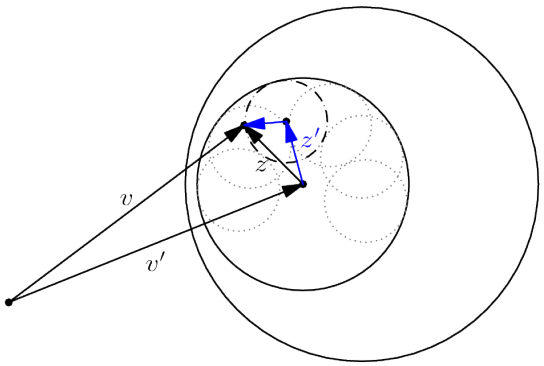

Figure 8.1: Illustration of a chained cover. Within the ϵ-ball containing the discretization error z, we find a finer ϵ'-cover and obtain a smaller error z' from discretizing z.

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

Theorem 8.14. *(Dudley chaining) Let $\mathcal{F}$ be a family of functions from Z to $\mathbb{R}$. Then*

$$R_S(\mathcal{F}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(p_n))}{n}} d\epsilon. \tag{8.54}$$

We can interpret this bound as removing the discretization error term by averaging over different scales of ϵ. For a proof of this theorem, refer to Theorem 15 of [Lia16].

Bibliography

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

[WLLM20] Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma, *Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel*, 2020.