FML Project: Generalization bound for neural network

David Huang SH6362@NYU.EDU
Devin Tang WT2113@NYU.EDU
New York University

Abstract

In this survey paper, we will investigate into two paper: 1) Spectrally-normalized margin bounds for neural networks. 2)The Sample Complexity of One-Hidden-Layer Neural Networks. In the first paper, we are considering margin-based multiclass generalization bound for neural networks that scaling with their margin-normalized spectral complexity–their Lipschitz constant. We first define the normalized margin and give the sketch to our main theorem on generalized bound. In the second paper, we are interested in what kind of norm control on hidden layer weight matrix will result in uniform convergence. Based on that, we also loose some of assumption in order to get more involved result on norm control of weight matrix

1. Spectrally-normalized margin bounds for neural networks

1.1. key observation

When we consider the measurable of complexity, we found an intriguing phenomena: even though complexity of standard networks exhibit growing Lipschitz constants, normalizing these Lipschitz constants by the margin instead gives a constant/decaying curve (i.e Lipchitz is increasing, but Lipchitz/margin is not)

1.2. unnormalized margin

Consider $f : \mathbb{R}^d \to \mathbb{R}^k$, where k is the number of classes; Intuitively, we define the classifier as selecting the output coordinate with the largest magnitude,

$$x \mapsto \text{argmax}_j f(x)_j$$

unnormalized margin is defined as $f(x)_y - \max_{j \neq y} f(x)_j$, which measures the gap between the output for the correct label and other labels (a measure of confidence)

Recall that unnormalized margin alone doesn't say too much as we discussed in the previous section, we need proper normalization.

1.3. normalized margin

Before we define the normalized margin, we need to first clarify some notation. Suppose the networks will use L fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$, where $\sigma_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ is ρ_i-Lipschitz with $\sigma_i(0) = 0$. Given L weight matrices $A = (A_1, \ldots, A_L)$. Then, we define the function

$$F_A(x) := \sigma_L(A_L \sigma_{L-1}(A_{L-1} \cdots \sigma_1(A_1 x) \cdots))$$

The classifier in this case is then naturally defined as:

$$x \mapsto \text{argmax}_j F_A(x)_j$$

Next, define a collection of reference matrices $(M_1, \ldots, M_L)$ with the same dimensions as $A_1, \ldots, A_L$. Let $\|\cdot\|_{p,q}$ denote the (p,q) matrix norm, defined by $\|A\|_{p,q} := \|(\|A_{:,1}\|_p, \ldots, \|A_{:,m}\|_p)\|_q$ for $A \in \mathbb{R}^{d\times m}$

The **spectral complexity** $R_{F_A} = R_A$ of a network F_A with weights A is the defined as

$$R_A := \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\right) \left(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\right)^{3/2}. \tag{1}$$

normalized margin is then defined as

$$\frac{F_A(x)_y - \max_{i\neq y} F_A(x)_i}{R_A\|X\|_2/n},$$

1.4. main theorem

Theorem 1 *Let $(x,y), (x_1,y_1), \ldots, (x_n,y_n)$ drawn iid from any probability distribution over $\mathbb{R}^d \times \{1,\ldots,k\}$, with probability at least $1-\delta$ over $((x_i,y_i))_{i=1}^n$, every margin $\gamma > 0$ and network $F_A : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $A = (A_1,\ldots,A_L)$ satisfy*

$$\mathbb{P}\left[\arg\max_j F_A(x)_j \neq y\right] \leq \widehat{R}_\gamma(F_A) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_2 R_A}{\gamma n}\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right)$$

where $\widehat{\mathcal{R}}_\gamma(f) \leq n^{-1}\sum_i \mathbb{1}\left[f(x_i)_{y_i} \leq \gamma + \max_{j\neq y_i} f(x_i)_j\right]$ and $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

What this theorem teaches us is following:
1) divide the Lipchitz constant ρ_i by the margin γ.2)no dependence on the combinatorical parameters. 3) no dependency on number of class. 4)measure complexity again reference network.

1.5. proof sketch

1.5.1. STEP 1:STANDARD RADEMACHER COMPLEXITY BOUND

By same argument in the class/textbook, we apply standard tool in Rademacher complexity:

$$\Pr\left[\arg\max_i f(x)_i \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(f) + \underbrace{2\mathcal{R}((\mathcal{F}_\gamma)_{|S})}_{*} + 3\sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{2}$$

To control $(*)$, we need to invoke the concept of covering numbers. $\mathcal{N}(U,\epsilon,\|\cdot\|)$ denote the least cardinality of any subset $V \subseteq U$ that covers U at scale ϵ with norm $\|\cdot\|$

1.5.2. STEP2: MATRIX COVERING(VIA MAUREY SPARSIFICATION LEMMA)

We first bound on each layer. Let conjugate exponents (p,q) and (r,s) be given with $p \leq 2$, as well as positive reals (a,b,ϵ) and positive integer m. X is the data passed through all layers prior to the present layer. $X \in \mathbb{R}^{n\times d}$ be given with $\|X\|_p \leq b$. Then

$$\ln\mathcal{N}\left(\left\{XA : A \in \mathbb{R}^{d\times m}, \|A\|_{q,s} \leq a\right\}, \epsilon, \|\cdot\|_2\right) \leq \left\lceil \frac{a^2b^2m^{2/r}}{\epsilon^2} \right\rceil \ln(2dm).$$

1.5.3. STEP3: INDUCTION TO THE WHOLE NETWORK

Let X_i denote the output of layer i, and $\widehat{X}_i$ is the covering element which depends on covering matrices $(\widehat{A}_1, \ldots, \widehat{A}_{i-1})$ chosen to cover weight matrices in prvious layer. Via matrix covering, $X_{i+1} = \sigma(A_i X_{i+1})$ with $\widehat{X}_{i+1} := \sigma(\widehat{A}_i \widehat{X}_i)$:. Then we have this relationship to apply induction

$$\|X_{i+1} - \widehat{X}_{i+1}\|_2 \leq \rho_i \|A_i\|_\sigma \|X_i - \widehat{X}_i\|_2 + \rho_i \epsilon_i,$$

1.5.4. STEP4:FULL NETWORK COVERING BOUNDS

Let $\mathcal{F}$ denote networks $F_{\mathcal{A}}$ where matrices $\mathcal{A} = (A_1, \ldots, A_L)$ satisfy $\|A_i\|_\sigma \leq s_i$, $\|A_i\|_{2,1} \leq b_i$; σ_i is ρ_i-Lipschitz with $\sigma_i(0) = 0$. Combining above pieces gives

$$\ln \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \leq \frac{\|X\|_2^2 \ln(2W^2)}{\epsilon^2} \left(\prod_{j=1}^{L} s_j^2 \rho_j^2\right) \left(\sum_{i=1}^{L} \frac{b_i}{s_i}\right)^{2/3}.$$

1.5.5. STEP5: DUDLEY ENTROPY INTEGRAL AND UNION BOUND

We plug in previous result in the following Dudley Entropy Integral bound:

$$\mathcal{R}(\mathcal{F}_{|S}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_S, \epsilon, \|\cdot\|_2)}\, d\epsilon \right).$$

main theorem is thus followed by cutting up parameter space and applying union bound

1.6. extension

Potential improvements to the main theorem: Are these choices of norms the best? We may want to choose a different norm from the spectral norm or choose them to be layer-specific. Will the choice of reference matrices affect the complexity bounds? Can the Lipschitz constant of the nonlinearities be replaced by a better quantity? These are questions worth our future research.

2. The Sample Complexity of One-Hidden-Layer Neural Networks

2.1. problem formulation

We are interested in norm-based uniform convergence bound for neural network, and we want to study what types of constraints on the network weights W can lead to uniform convergence bounds. Consider the neural network from $\mathbb{R}^d \to \mathbb{R}$ of the following form:

$$\mathbf{x} \mapsto \mathbf{u}^\top \sigma(\mathbf{W}\mathbf{x})$$

where $W \in \mathbb{R}^{n \times d}$ be the weight matrix, $\mathbf{u} \in \mathbb{R}^n$ be the weight vector, and σ is an activation function.

2.2. fat-shattering dimension and Rademacher complexity

In order to **lower** bound sample complexity of a given function class, we invoke fat-shattering dimension:

A class of functions $\mathcal{F}$ on an input domain $\mathcal{X}$ shatters m points $\{\mathbf{x}_i\}_{i=1}^m \subseteq \mathcal{X}$ with margin ϵ, if there exist a number s, such that for all $\mathbf{y} \in \{0,1\}^m$, we can find some $f \in \mathcal{F}$ such that

$$\forall i \in [m], \quad f(\mathbf{x}_i) \leq s - \epsilon \text{ if } y_i = 0 \quad \text{and} \quad f(\mathbf{x}_i) \geq s + \epsilon \text{ if } y_i = 1.$$

The fat-shattering dimension of $\mathcal{F}$ (at scale ϵ) is the cardinality m of the largest set of points in $\mathcal{X}$ for which the above holds.

To **upper** bound sample complexity, we invoke Rademacher complexity defined as follows:

$$\mathcal{R}_m(\mathcal{F}) = \sup_{\{x_i\}_{i=1}^m \subseteq \mathcal{X}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \epsilon_i f(x_i) \right]$$

2.3. spectral norm is insufficient(dimension free)

In this section, we have no assumption on input dimension d (**dimension free**) in the sense that upper bound for any d and lower bound hold for large d.

Consider the following hypothesis class:

$$\mathcal{H}_{b,B,n,d}^{\sigma} := \left\{ \mathbf{x} \mapsto \mathbf{u}^\top \sigma(W\mathbf{x}) : \mathbf{u} \in \mathbb{R}^n, W \in \mathbb{R}^{n \times d}, \|\mathbf{u}\| \leq b, \|W\| \leq B \right\}$$

Note that the constraint on u,W is of spectral norm. We obtain the result that:

if input dimension is large enough, mild condition on σ's non-smoothness in the sense that suppose that the activation function σ is 1-Lipschitz on $[-1,+1]$, and satisfies $\sigma(0) = 0$ and for some $\alpha > 0$ as well as $\inf_{\delta \in (0,1)} \left| \frac{\sigma(\delta)+\sigma(-\delta)}{\delta} \right| \geq \alpha$.. The fat-shattering dimension of the class scales with network width n. In another word, we never get bounds independent of the width n

In conclusion, the covering number is lower bounded as $\Omega(\frac{B^2 n}{\epsilon^2})$

2.4. Frobenius norm is sufficient(dimension free)

We start with basic fact: if M is an $n \times d$ matrix, then $\|M\|_F \leq \|M\| \cdot \sqrt{\min\{n,d\}}$.
In our case, $B^2 n$ is thus the upper bound for $\|W\|_F^2$. Therefore , in order to get the width independent bound, we need to directly controll on $\|W\|_F$. Then we consider the following hyphethesis class:

$$\mathcal{F}_{b,B,n,d}^{\sigma} := \left\{ \mathbf{x} \mapsto \mathbf{u}^\top \sigma(W\mathbf{x}) : \mathbf{u} \in \mathbb{R}^n, W \in \mathbb{R}^{n \times d}, \|\mathbf{u}\| \leq b, \|W\|_F \leq B \right\}$$

Theorem 2 *Suppose $\sigma(\cdot)$ (as a function on $\mathbb{R}$) is L-Lipschitz and $\sigma(0) = 0$. Then for any $b, B, b_x, n, d, \epsilon > 0$, the Rademacher complexity of $\mathcal{F}_{b,B,n,d}^{\sigma}$ on m inputs from $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq b_x\}$ is at most ϵ, if*

$$m \geq c \cdot \frac{(bBb_xL)^2(1 + \log^3(m))}{\epsilon^2},$$

for some universal constant $c > 0$. Thus, it suffices to have $m = \widetilde{\mathcal{O}}\left(\frac{(bBb_xL)^2}{\epsilon^2}\right)$.

Thus, we get the sample complexity upper bound $\mathcal{O}\left(\frac{B^2}{\epsilon^2}\right)$, where B bounds Frobenius norm.

2.5. dimension dependent lower bound

We should notice that when input dimension d is fixed, although theorem 2 is bounded on Rademacher complexity, fat-shattering dimension could be smaller for dimension d.

For the Frobenius norm case, we can have following argument regarding dimension dependent lower bound for a ReLU activation function with a bias term β: $\widetilde{\Omega}\left(\min\left\{nd, \frac{bBb_x}{\epsilon}\sqrt{d}\right\}\right)$

2.6. when spectral control work?

2.6.1. SMOOTHNESS ASSUMPTION

Previous analysis based on the fact that activation function σ is non-smooth. We also know that spectral norm control is never sufficient. A natural question is that if we change our assumption regarding smoothness of σ, can we say more? In this section, we will present the result that indeed for sufficiently smooth activation(i.e $poly(k)$, $erf(\cdot)$), even we only have control on the spectral norm, we can surprisingly get a width-independent Rademacher complexity bound.

In another word, for $\sigma(z) = \sum_{j=1}^{\infty} a_j z^j$ and $\widetilde{\sigma}(z) = \sum_{j=1}^{\infty} |a_j| z^j$, we have sample complexity $O\left(\frac{\widetilde{\sigma}(B)^2}{\epsilon^2}\right)$. Here is some example with different activation function:

	$\sigma(z)$	$\widetilde{\sigma}(B_x)$
polynomial of degree k	poly(k)	$\mathcal{O}((Bb_x)^k)$
error function	$erf(rz)$	$\mathcal{O}(\frac{2rBb_x}{\sqrt{\pi}} \exp\left((rBb_x)^2\right))$
sigmoidal	$\frac{1}{2}\left(y + \int_{z=0}^{y} \mathrm{erf}(rz)\, dz\right)$	$\mathcal{O}(\frac{Bb_x}{2} + \frac{r(Bb_x)^2}{\sqrt{\pi}} \exp\left((rBb_x)^2\right))$

2.7. convolutional network

In this section, in order to circumvent the dimension dependent lower bound seen in section 2.5, we now consider hidden-layer convolutional neural networks. Consider the patches $\Phi = \{\phi_j\}_{j=1}^{n}$. For each j, ϕ_j project input vector into $x \in \mathbb{R}^d$ into some subset of coordinate. the output vector can be written as $\sigma(Wx)$. Under assumption of conformality, $(Wx)_j = w^T \phi_j(x)$.

The result can be summarized in the following form:

	output layer assumption	**sample comlexity**
linear output layer	$u^T \sigma(Wx)$	$2 \cdot O_\Phi \cdot \left(\frac{bBb_xL}{\epsilon}\right)^2$
pooling layer	$\rho \circ \sigma(W\mathbf{x})$	$c \cdot \left(\frac{LBb_x}{\epsilon}\right)^2 \cdot \log^2(m)\log(mn)$

where $\mathbf{x} \mapsto \rho \circ \sigma(W\mathbf{x}) = \rho\left(\sigma(\mathbf{w}^\top \phi_1(\mathbf{x})), \ldots, \sigma(\mathbf{w}^\top \phi_n(\mathbf{x}))\right)$. Note in linear output layer, u^T is a vector in $\mathbb{R}^n$, and in the pooling layer case, where $\rho : \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz with respect to the ℓ_∞ norm. $\rho(\cdot)$ may correspond to a max-pooling layer $\mathbf{z} \mapsto \max_{j\in[n]} z_j$, or to an average-pooling layer $\mathbf{z} \mapsto \frac{1}{n}\sum_{j\in[n]} z_j$.

2.8. extension

A very interesting question for me to consider in the further is that: Recall we have observed how smoothness of activation function σ can play a crucial role in the in spectral norm controlled case. However, we still do not obtain enough understanding on when we can get width-gree gurantee in the spectral norm case. I think it will be very interesting to systematically establish a class for the case that can lead to width-free guanrantee. Or another question: can we change our other assumption to make this guarantee not dependent on smoothness?

References

[1] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] Gal Vardi, Ohad Shamir, and Nathan Srebro. The sample complexity of one-hidden-layer neural networks. In *Proceedings of the 34th Annual Conference on Learning Theory (COLT)*, 2021.