Covering and Rademacher bounds for neural networks

Overview

This lecture is the first of three lectures investigating the complexity of neural networks. We'll cover 3 bounds, 2 of which use covering numbers, and at the end point show how to relate them to Rademacher complexity.

The definition of cover we'll use is as follows.

Definition. Say G is a $(\|\cdot\|_p; \epsilon, S)$-cover of $\mathcal{F}$ if:

- For every $f \in \mathcal{F}$, there exists $g \in G$ so that $\|g(S) - f(S)\|_p \leq \epsilon$, where $g(S) := (g(x_1), \ldots, g(x_n))$ and $S := (x_1, \ldots, x_n)$.
- $\sup_{g \in G} \|g(S)\|_2 \leq 2 \sup_{f \in \mathcal{F}} \|f(S)\|_2$. (We need this to apply Rademacher rules for finite classes.)

Moreover, define the **covering number** $\mathcal{N}_p(\mathcal{F}; \epsilon, S)$ to be minimal cardinality of a $(\|\cdot\|_p; \epsilon, S)$-cover of $\mathcal{F}$ (or ∞ if no finite cover exists).

The three bounds are as follows.

1. By ignoring the structure of networks and treating them as Lipschitz functions, the Rademacher complexity can be upper bounded by a quantity exponential in layers and dimension.
2. By paying a little attention to the structure, a bound which is exponential only in layers can be derived.
3. Lastly, if more attention is paid and the activations are bounded, then the complexity is polynomial in layers and parameters.

Remark.

- Networks here are of the form $\vec{\phi}(A_l \vec{\phi}(A_{l-1} \vec{\phi}(\cdots A_1 x)))$, and in particular the offset "$+b$" is omitted".
- The covers here are *improper*, meaning we do not require $G \subseteq \mathcal{F}$.
- The notion of covers (including improper covers) is due to Kolmogorov and Tikhomorov *[future matus: read it ...]*

Pure lipschitz bound

The first bound is simply on classes of functions which satisfy a certain Lipschitz constant.

Theorem. Let $S = (x_1, \ldots, x_n)$ be given with $R := \max_{i,j} \|x_i - x_j\|_\infty$. Let $\mathcal{F}$ denote the collection of all L-lipschitz functions that have range $[-B, +B]$ for some $B \geq 0$. Then

$$\ln \mathcal{N}_\infty(\mathcal{F}; \epsilon, S) \leq \max \left\{ 0, \left\lceil \frac{2LR}{\epsilon} \right\rceil^d \ln \left\lceil \frac{2B}{L\epsilon} \right\rceil \right\}.$$

Proof. If $B \leq \epsilon$, then $G = \{x \mapsto 0\}$ suffices as a cover. Otherwise suppose $B > \epsilon$, and take G to be the following class of piecewise-constant functions (meaning the elements of G are in general not Lipschitz). Let U denote any cube of side length R which contains S. Subdivide U into cubes of side length $\epsilon/(2L)$, which

yields a partition P of U into $\lceil 2LR/\epsilon \rceil^d$ cubes. Next, discretize $[-B, +B]$ at resolution ϵ, which yields a partition Q of size $\lceil 2B/\epsilon \rceil$. The cover G is now all possible mappings from P to Q, and thus

$$\ln |G| \leq \ln |Q|^{|P|} = \lceil 2LR/\epsilon \rceil^d \ln \lceil 2B/\epsilon \rceil.$$

To verify that G covers $\mathcal{F}$, first note that $\sup_{g \in G} g(S) = B\sqrt{n} = \sup_{f \in \mathcal{F}} \|f(S)\|_2$ by considering the function which is equal to $-B$ everywhere, which is within $G \cap \mathcal{F}$. Moreover, given any $f \in \mathcal{F}$, select from g the function which is as close to f as possible at the midpoint of any cube in the partition P. Consequently, for any $x \in S$, letting x' denote the midpoint of the cube containing x,

$$|g(x) - f(x)| \leq |g(x) - g(x')| + |g(x') - f(x')| + |f(x') - f(x)| \leq 0 + \frac{\epsilon}{2} + L\|x - x'\|_\infty \leq \frac{\epsilon}{2} + L\left(\frac{\epsilon}{2L}\right).$$

□

To connect this to neural nets, it suffices to compute their Lipschitz constant.

Corollary *(see also proof of Lemma 14.6 in Anthony-Bartlett.)* **.** Suppose a neural network with k layers where ϕ is L-lipschitz and each node computes $z \mapsto \phi(a^\top z)$ for some parameter vector a with $\|a\|_1 \leq W$. Then this network has Lipschitz constant at most $(LW)^k$, and moreover given a sample S which lies in the l_∞ ball at the origin of diameter R, the class $\mathcal{F}$ of all such networks satisfies

$$\ln \mathcal{N}_\infty(\mathcal{F}; \epsilon, S) \leq \max\left\{0, \left\lceil \frac{2R(LW)^k}{\epsilon} \right\rceil^d \ln \left\lceil \frac{2R}{\epsilon} \right\rceil \right\}.$$

Proof. Note that the covering number follows from the Lipschitz estimate. In turn, the Lipschitz bound will be proved by induction on layers; namely, it will be shown that every node in layer i Lipschitz constant at most $(LW)^i$.

The base case is layer $i = 0$ of the inputs themselves, which have Lipschitz constant $1 = (LW)^i$: for any coordinate $i \in [d]$ and any inputs x, x',

$$|x_i - x'_i| \leq 1 \cdot \|x - x'\|_\infty.$$

In case of layer $i + 1$ with $i \geq 0$, let $H(x)$ denote the output of layer i, whereby the output of any node in layer $i + 1$ with parameter vector a satisfies for any x, x' with

$$\begin{aligned} |\phi(a^\top H(x)) - \phi(a^\top H(x'))| &\leq L|a^\top H(x)) - a^\top H(x')| \\ &\leq L\|a\|_1 \|H(x) - H(x')\|_\infty \\ &\leq LW \max_j |H(x)_j - H(x')_j| \\ &\leq (LW)^{i+1}. \end{aligned}$$

□

Direct Rademacher analysis of Lipschitz neural networks

The previous bounds were insensitive to the structure of neural networks, but were instead bounds on Lipschitz functions. By studying neural networks directly, it is possible to produce a bound which removes the exponential dependence on dimension.

Theorem *(bound implied by Anthony-Bartlett Theorem 14.5 proof; similar bound in Karthik's notes)* **.** Consider a network with k layers where nodes compute $z \mapsto \phi(a^\top z)$ where parameters a varies by node and $\|a\|_1 \leq W$ and ϕ is L-lipschitz with $\phi(0) = 0$. Lastly suppose each $x \in S$ has $\|x\|_\infty \leq X$. Then

$$\mathrm{Rad}((\mathcal{F}_i)_{|S}) \leq X(LW)^k \sqrt{\frac{2\ln(2d)}{n}}.$$

Proof. Let $\mathcal{F}_i$ denote the functions computed by nodes in layer i. It will be shown by induction that $\sup_{f\in\mathcal{F}_i}\langle\sigma, f(S)\rangle \geq 0$ for every $\sigma\in\{-1,+1\}^n$ and

$$\mathrm{Rad}((\mathcal{F}_i)_{|S}) \leq X(LW)^i\sqrt{\frac{2\ln(2d)}{n}}.$$

In the base case $i=0$, take $\mathcal{F}_0$ to consist of the d coordinate maps as well as their negations. Including the negations means $\mathcal{F}_0 = -\mathcal{F}_0$ and thus the nonnegativity condition holds, and moreover the finite class lemma gives

$$\mathrm{Rad}((\mathcal{F}_0)_{|S}) \leq \frac{\max_{i\in d}\|(x_1)_i,\ldots,(x_n)_i\|_2\sqrt{2\ln(2d)}}{n} \leq X(2LW)^0\sqrt{\frac{2\ln(2d)}{n}}.$$

In the case of layer $i+1$ with $i\geq 0$, the nonnegativity condition holds since $a=0$ is possible and $\phi(0)=0$, and moreover

$$\begin{aligned}
\mathrm{Rad}((\mathcal{F}_{i+1})_{|S}) &\leq L\mathrm{Rad}((W\mathrm{conv}(\mathcal{F}_i\cup-\mathcal{F}_i))_{|S})\\
&\leq LW\mathrm{Rad}(\mathrm{conv}(\mathcal{F}_i\cup-\mathcal{F}_i)_{|S})\\
&\leq LW\mathrm{Rad}((\mathcal{F}_i\cup-\mathcal{F}_i)_{|S})\\
&\leq 2LW\mathrm{Rad}((\mathcal{F}_i)_{|S})\\
&\leq X(2LW)^{i+1}\sqrt{\frac{2\ln(2d)}{n}}.
\end{aligned}$$

□

Covering number bound for neural networks with bounded nonlinearities

Theorem *(simplification of Theorem 14.5 in Anthony-Bartlett)* **.** Suppose $\phi:\mathbb{R}\to[-B,+B]$ is L-lipschitz; suppose the networks have p parameters, k layers, and $\leq m$ nodes per layer; suppose the input dimension d satisfies $d\leq m$; suppose each node computes $z\mapsto\phi(a^\top z)$ where $\|a\|_1\leq W$; lastly suppose $L\geq 1$ and $W\geq 1$. Then for any sample $S\subseteq[-B,+B]^d$, letting $\mathcal{F}$ denote all networks satisfying the above conditions,

$$\mathcal{N}_\infty(\mathcal{F};\epsilon,S) \leq \left\lceil\frac{2mWB(2LW)^k}{\epsilon}\right\rceil^p.$$

Proof. The cover is constructed as follows. Suppose each of the p parameters are discretized along $[-W,+W]$ at scale ϵ_0/m (where ϵ_0 will be determined later), whereby the size of the final cover is $\lceil 2mW/\epsilon_0\rceil^p$. By this choice, the parameter vector a in any node has a discretized approximant $\hat{a}$ which satisfies

$$\|a-\hat{a}\|_1 \leq \sum_{i=1}^{\leq m}|a_i-\hat{a}_i| \leq m\epsilon_0/m = \epsilon_0.$$

Now let $\mathcal{F}_i$ denote the class of functions computed by nodes of layer i, meaning $\mathcal{F}=\mathcal{F}_k$. It will be proved by induction that for every function g compute by a node in layer i, there exists a function $\hat{g}$ obtained by discretizing the weights of g as above so that every $x\in S$ satisfies

$$|g(x)-\hat{g}(x)| \leq B\epsilon_0(2LW)^i.$$

This suffices to prove the theorem by choosing $\epsilon_0 := \epsilon/(B(2LW)^k)$ and recalling the cover has size $\lceil 2mW/\epsilon_0\rceil^p$.

To establish the induction, first note that layer 0 depends on no parameters and is represented perfectly. For layer $i+1$ with $i\geq 0$, consider some node computing $g(x)=\phi(a^\top H(x))$ where a is a parameter vector and H represents the outputs of the previous layer. The approximant $\hat{g}$ can be written as $\hat{g}(x)=\phi(\hat{a}^\top\hat{H}(x))$

where $\hat{a}$ is the discretization of a and $\hat{H}$ is obtained from H by discretizing all the weights in previous layers. By the inductive hypothesis,

$$\begin{aligned}
|g(x) - \hat{g}(x)| &= |\phi(a^\top H(x)) - \phi(\hat{a}^\top \hat{H}(x))| \\
&\leq L|a^\top H(x) - \hat{a}^\top \hat{H}(x)| \\
&\leq L|a^\top H(x) - a^\top \hat{H}(x)| + L|a^\top \hat{H}(x) - \hat{a}^\top \hat{H}(x)| \\
&\leq L\|a\|_1 \|H(x) - \hat{H}(x)\|_\infty + L\|a - \hat{a}\|_1 \|\hat{H}(x)\|_\infty \\
&\leq LW(B\epsilon_0(2LW)^i) + L\epsilon_0 B = \epsilon_0 B(LW(2LW)^i + L) \leq \epsilon_0 B(2LW)^{i+1}.
\end{aligned}$$

□

Covering and Rademacher

To close, note how cover and Rademacher complexities may be compared; note that this bound is somewhat primitive and a tighter way will be covered in homework. Also, this proof is similar to the direct union bound method we had earlier of converting covers into generalization bounds.

Theorem.

$$\mathrm{Rad}(\mathcal{F}_{|S}) \leq \inf_{\epsilon > 0} \left(\frac{\epsilon}{n^{1/p}} + \frac{\sup_{f \in \mathcal{F}} 2\|f(S)\|_2 \sqrt{2\ln(\mathcal{N}_p(\mathcal{F}; \epsilon, S))}}{n} \right).$$

Proof. Let $\epsilon > 0$ be arbitrary, let G be a $(\|\cdot\|_p; \epsilon, S)$-cover of $\mathcal{F}$, and for any $f \in \mathcal{F}$ let $g_f \in G$ denote a closest approximant. Taking q to be the conjugate exponent of p (meaning $1/p + 1/q = 1$),

$$\begin{aligned}
\mathrm{Rad}(\mathcal{F}_{|S}) &= \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left\langle \sigma, f(S) \right\rangle / n \\
&= \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left(\left\langle \sigma, g_f(S) \right\rangle / n + \left\langle \sigma, f(S) - g_f(S) \right\rangle / n \right) \\
&\leq \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left(\left\langle \sigma, g_f(S) \right\rangle / n + \|\sigma\|_q \|f(S) - g_f(S)\|_p / n \right) \\
&\leq \mathrm{Rad}(G_{|S}) + \epsilon n^{1/q - 1} \\
&\leq \epsilon n^{-1/p} + \frac{\sup_{g \in G} \|g(S)\|_2 \sqrt{2\ln(|G|)}}{n} \\
&\leq \epsilon n^{-1/p} + \frac{\sup_{f \in \mathcal{F}} 2\|f(S)\|_2 \sqrt{2\ln(\mathcal{N}_p(\mathcal{F}; \epsilon, S))}}{n}.
\end{aligned}$$

□

References