Generalization errors for Deep Learning: Going Deep

김지수 (Jisu KIM)

딥러닝의 통계적 이해 (Deep Learning: Statistical Perspective), 2024년 2학기

This lecture note is a combination of Prof. Joong-Ho Won's "Deep Learning: Statistical Perspective" with other lecture notes. Main references are:

Tong Zhang, Mathematical Analysis of Machine Learning Algorithms, https://tongzhang-ml.org/lt-book.html

Matus Telgarsky, Deep learning theory lecture notes, https://mjt.cs.illinois.edu/dlt/

Weinan E, Chao Ma, Stephan Wojtowytsch, Lei Wu, Towards a Mathematical Understanding of Neural Network-Based Machine Learning: what we know and what we don't, https://arxiv.org/abs/2009.10713/

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^d$, so $x = (x_1, \ldots, x_d)$.
- Output(출력) / Response(반응 변수): $y \in \mathcal{Y}$. If y is categorical, then supervised learning is "classification", and if y is continuous, then supervised learning is "regression".
- Model(모형) :
$$y \approx f(x).$$
If we include the error ϵ to the model, then it can be also written as
$$y = \phi(f(x), \epsilon).$$
For many cases, we assume additive noise, so
$$y = f(x) + \epsilon.$$
- Assumption(가정): f belongs to a family of functions $\mathcal{M}$. This is the assumption of a model: a model can be still used when the corresponding assumption is not satisfied in your data.
- Loss function(손실 함수): $\ell(y, a)$. A loss function measures the difference between estimated and true values for an instance of data.
- Training data(학습 자료): $\mathcal{T} = \{(y_i, x_i), i = 1, \ldots, n\}$, where (y_i, x_i) is a sample from a probability distribution P_i. For many cases we assume i.i.d., or x_i's are fixed and y_i's are i.i.d..
- Goal(목적): we want to find f that minimizes the expected prediction error,
$$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$
Here, $\mathcal{F}$ can be different from $\mathcal{M}$; $\mathcal{F}$ can be smaller then $\mathcal{M}$.
- Prediction model(예측 모형): f^0 is unknown, so we estimate f^0 by $\hat{f}$ using data. For many cases we minimizes on the empirical prediction error, that is taking the expectation on the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i)}$.
$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n} \left[\ell(Y, f(X))\right] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$
- Prediction(예측): if $\hat{f}$ is a predicted function, and x is a new input, then we predict unknown y by $\hat{f}(x)$.

1.2 Rademacher complexity

Random variables $\xi_1, \ldots, \xi_n$ are called *Rademacher random variables* if they are independent, identically distributed and $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1/2$. Define the *Rademacher complexity* of $\mathcal{F}$ by

$$\mathsf{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \xi_i f(Z_i)\right)\right).$$

Some authors use a slightly different definition, namely,

$$\mathsf{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(Z_i)\right|\right).$$

You can use either one. They lead to essentially the same results. In fact, under mild condition, two Rademacher complexity are closely related, as below:

Lemma 1. *Let $\xi = (\xi_1, \ldots, \xi_n)$ be i.i.d. Rademacher. Suppose that for any $\xi \in \{\pm 1\}^n$, $\sup_{f\in\mathcal{F}} \sum_{i=1}^n \xi_i f(Z_i) \geq 0$. Then*

$$\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(Z_i)\right| \middle| Z\right] \leq 2\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \xi_i f(Z_i) \middle| Z\right].$$

Proof. Left as HW. □

Intuitively, $\mathsf{Rad}_n(\mathcal{F})$ is large if we can find functions $f \in \mathcal{F}$ that "look like" random noise, that is, they are highly correlated with $\sigma_1, \ldots, \sigma_n$. Here are some properties of the Rademacher complexityd

Lemma. *(a) If $\mathcal{F} \subset \mathcal{G}$ then $\mathsf{Rad}_n(\mathcal{F}, Z^n) \leq \mathsf{Rad}_n(\mathcal{G}, Z^n)$.*

(b) Let $\mathsf{conv}(\mathcal{F})$ denote the convex hull of $\mathcal{F}$. Then $\mathsf{Rad}_n(\mathcal{F}, Z^n) = \mathsf{Rad}_n(\mathsf{conv}(\mathcal{F}), Z^n)$.

(c) For any $c \in \mathbb{R}$, $\mathsf{Rad}_n(c\mathcal{F}, Z^n) = |c|\mathsf{Rad}_n(\mathcal{F}, Z^n)$.

(d) Let $g : \mathbb{R} \to \mathbb{R}$ be such that $|g(y) - g(x)| \leq L|x - y|$ for all x, y. Then $\mathsf{Rad}_n(g \circ \mathcal{F}, Z^n) \leq L\mathsf{Rad}_n(\mathcal{F}, Z^n)$.

(e) Suppose $\{\mathcal{F}_i\}_{i\in I}$ satisfies $0 \in \mathcal{F}_i$ for each $i \in I$. Then $\mathsf{Rad}_n(\bigcup_{i\in I}\mathcal{F}, Z^n) \leq \sum_{i\in I}\mathsf{Rad}_n(\mathcal{F}_i, Z^n)$.

1.3 Two Layer Neural Networks

A two-layer neural network takes an input vector of d variables $x = (x_1, x_2, \ldots, x_d)$ and builds a nonlinear function $f(x)$ to predict the response $y \in \mathbb{R}^D$. What distinguishes neural networks from other nonlinear methods is the particular structure of the model:

$$f(x) = f_\theta(x) = g\left(\beta_0 + \sum_{j=1}^m \beta_j \sigma(b_j + w_j^\top x)\right),$$

where $x \in \mathbb{R}^d, b_j \in \mathbb{R}, w_j \in \mathbb{R}^d, \beta_0 \in \mathbb{R}^D, \beta_j \in \mathbb{R}^D$. See Figure 1.

- $\theta = \{[\beta, a_j, b_j, w_j] : j = 1, \ldots, m\}$ denotes the set of model parameters.
- $x_1, \ldots, x_d$ together is called an input layer.
- $A_j \coloneqq \sigma_j(x) = \sigma(b_j + w_j^\top x)$ is called an activation.
- $A_1, \ldots, A_m$ together is called a hidden layer or hidden unit; m is the number of hidden nodes.
- $f(x)$ is called an output layer.
- g is an output function. Examples are:
  - softmax $g_i(x) = \exp(x_i)/\sum_{l=1}^D \exp(x_l)$ for classification. The softmax function estimates the conditional probability $g_i(x) = P(y = i|x)$.

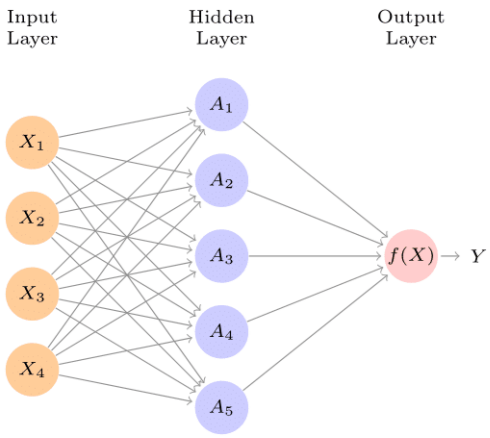

Figure 1: Neural network with a single hidden layer. The hidden layer computes activations $A_j = \sigma_j(x)$ that are nonlinear transformations of linear combinations of the inputs $x_1, \ldots, x_d$. Hence these A_j are not directly observed. The functions σ_j are not fixed in advance, but are learned during the training of the network. The output layer is a linear model that uses these activations A_j as inputs, resulting in a function $f(x)$. Figure 10.1 from [4].

  - identity/linear $g(x) = x$ for regression.
  - threshold $g_i(x) = I(x_i > 0)$

- σ is called an activation function. Examples are:
  - sigmoid $\sigma(x) = 1/(1 + e^{-x})$ (see Figure 2)
  - rectified linear (ReLU) $\sigma(x) = \max\{0, x\}$ (see Figure 2)
  - identity/linear $\sigma(x) = x$
  - threshold $\sigma(x) = I(x > 0)$, threshold gives a direct multi-layer extension of the perceptron (as considered by Rosenblatt).

Activation functions in hidden layers are typically nonlinear, otherwise the model collapses to a linear model. So the activations are like derived features - nonlinear transformations of linear combinations of the features.

1.4 Multi Layer Neural Networks

Modern neural networks typically have more than one hidden layer, and often many units per layer. In theory a single hidden layer with a large number of units has the ability to approximate most functions. However, the learning task of discovering a good solution is made much easier with multiple layers each of modest size.

A deep neural network refers to the model allowing to have more than 1 hidden layers: given input $x \in \mathbb{R}^d$ and response $y \in \mathbb{R}^D$, to predict the response y. K-layer fully connected deep neural network is to build a nonlinear function $f(x)$ as

- Let $m^{(0)} = d$ and $m^{(K)} = D$
- Define recursively

$$
\begin{aligned}
&x^{(0)} = x, \quad (x \in \mathbb{R}^{m^{(0)}}),\\
&x_j^{(k)} = \sigma(b_j^{(k)} + (w_j^{(k)})^\top x^{(k-1)}), \quad w_j^{(k)}, x^{(k-1)} \in \mathbb{R}^{m^{(k-1)}}, b_j \in \mathbb{R}^{m^{(k)}}, \qquad k = 1, \ldots, K.\\
&f(x) = g(x^{(K)}).
\end{aligned}
$$

- $\theta = \{[b_j^{(k)}, w_j^{(k)}] : k = 1, \ldots, K,\ j = 1, \ldots, m^{(k)}\}$ denotes the set of model parameters.
- $m^{(k)}$ is the number of hidden units at layer k.

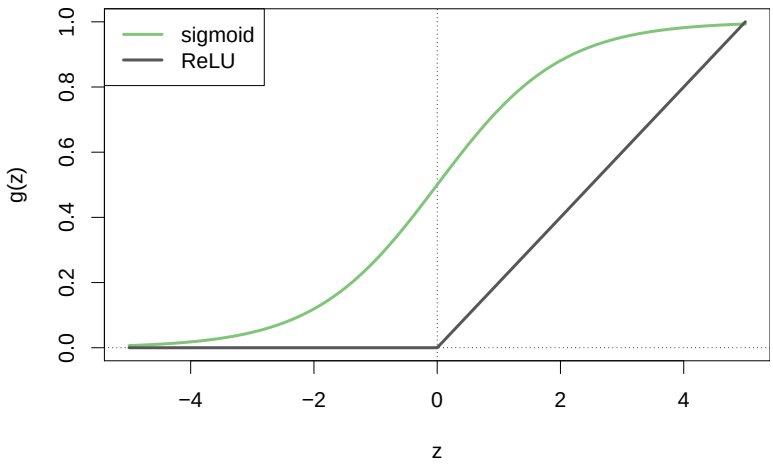

Figure 2: Activation functions. The piecewise-linear ReLU function is popular for its efficiency and computability. We have scaled it down by a factor of five for ease of comparison. Figure 10.2 from [4].

2 Notation and Goal

From here, we only consider regression problem, so $g(x) = x$. We assume $\beta_0 = 0$ and $b_j = 0$.

For the multi-layer neural network, for $k = 1, \dots, K$, write $W_k \in \mathbb{R}^{m^{(k)} \times m^{(k-1)}}$ as i-th row of W_k is $(w_j^{(k)})^\top$, i.e.,

$$W_k = \begin{pmatrix} (w_1^{(k)})^\top \\ \vdots \\ (w_{m^{(k)}}^{(k)})^\top \end{pmatrix} \in \mathbb{R}^{m^{(k)} \times m^{(k-1)}},$$

And for $k = 1, \dots, K-1$, write $\sigma_k : \mathbb{R}^{m^{(k)}} \to \mathbb{R}^{m^{(k)}}$ be a multivariate activation function. For example, for a univariate activation function σ, σ_k can be set as $\sigma_k(x_1, \dots, x_{m^{(k)}}) = (\sigma(x_1), \dots, \sigma(x_{m^{(k)}}))$. And let $\sigma_K \coloneqq g$. Now, assume $b_j^{(k)} = 0$ for $k = 1, \dots, K$, and $g = \mathrm{id}$. Then K-layer neural network can be described as

$$f_\theta(x) = \sigma_K(W_K \sigma_{K-1}(W_{K-1} \cdots \sigma_1(W_1 x) \cdots)).$$

Or inductively,

$$f_\theta^{(0)}(x) = x, \qquad f_\theta^{(k)}(x) = \sigma_k(W_k f_\theta^{(k-1)}(x)), \qquad f_\theta(x) = f_\theta^K(x).$$

We impose the condition that $\|W_k\| \le B$ for each k, where $\|\cdot\|$ is an appropriate matrix norm. Hence, for K-layer neural network, the function space we consider is $\mathcal{F}_\sigma^{(K)}$, with $\mathcal{F}_\sigma^{(0)} = \{\mathrm{id}\}$ and

$$\mathcal{F}_\sigma^{(k)} = \left\{ f_\theta^{(k)} : f_\theta^{(k)}(x) = \sigma_k(W_k f_\theta^{(k-1)}(x)),\ f_\theta^{(k-1)} \in \mathcal{F}_\sigma^{(k-1)}, \|W_k\| \le B \right\}.$$

Suppose the true regression function f_* is in a function class $\mathcal{M}$, so

$$y \approx f_*(x), \qquad f_* \in \mathcal{M}.$$

Suppose are using the ℓ_2-loss, so we find f among deep neural network class $\mathcal{F}$ that minimizes the expected risk (평균위험),

$$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X) \sim P} \left[(y - f(x))^2 \right].$$

f_0 is the expected risk mimizing function (평균위험최소함수). And we estimate f^0 by $\hat{f}$ using data by minimizes on the empirical risk (경험위험) on training dataset, so

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

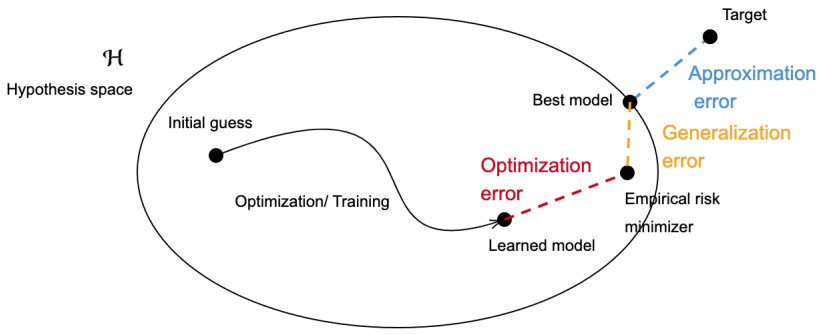

Figure 3: Diagram representing the learning procedure, the three main paradigms and their corresponding errors. Figure 2 from https://dcn.nat.fau.eu/breaking-the-curse-of-dimensionality-with-barron-spaces/.

$\hat{f}$ is the empirical risk mimizing function (경험위험최소함수). And we set $\tilde{f}$ be the approximation of $\hat{f}$ by optimization(최적화); $\tilde{f}$ is the learned function (학습된 함수).

So there are three sources of errors: approximation error, generalization error, and optimization error. See Figure 3.

$$f_* - \tilde{f} = \underbrace{f_* - f^0}_{\text{approximation error}} + \underbrace{f^0 - \hat{f}}_{\text{generalization error}} + \underbrace{\hat{f} - \tilde{f}}_{\text{optimization error}} .$$

3 Matrix norm, Sampling in Hilbert spaces, and basic Rademacher-covering relationship

For a matrix M, the norm $\|M\|_{b,c}$ for $b, c \geq 1$ means that, by letting $M^{(j)}$ be j-th column of M, then first apply b-norm to columns, and then c-norm to resulting vector, i.e.,

$$\|M\|_{b,c} = \left(\sum_j \left\| M^{(j)} \right\|_b^c \right)^{1/c} = \left(\sum_j \left(\sum_i |M_{ij}| \right)^{c/b} \right)^{1/c} .$$

Frobenius norm $\|M\|_F$ is $\|M\|_{2,2}$, i.e.

$$\|M\|_F = \left(\sum_{i,j} |M_{ij}|^2 \right)^{1/2} .$$

For $q \geq 1$, the operator norm is defined as

$$\|M\|_q = \sup_{x \neq 0} \frac{\|Mx\|_q}{\|x\|_q} = \sup_{x : \|x\|_q = 1} \|Mx\|_q .$$

Suppose $\mu = \mathbb{E}V$, where a random variable V is supported on a set S. A natural way to "simplify" μ is to instead consider $\bar{V} := \frac{1}{N} \sum_{i=1}^N V_i$, whre $(V_1, \ldots, V_N)$ are sampled i.i.d.. We want to argue $\bar{V} \approx \mu$; since we're considering a Hilbert space, we'll try to make the Hilbert norm $\left\| \mu - \bar{V} \right\|$ small.

Lemma 2. *Let $\mu = \mathbb{E}V$ be given, with V supported on S, and let $(V_1, \ldots, V_N)$ be i.i.d. draws from the same distribution. Then*

$$\mathbb{E}_{V_1,\ldots,V_N} \left\| \mu - \frac{1}{N} \sum_{i=1}^N V_i \right\|^2 \leq \frac{\mathbb{E} \|V\|^2}{N} \leq \frac{\sup_{U \in S} \|U\|^2}{N},$$

and moreover ther exist $(U_1, \ldots, U_N)$ in S so that

$$\left\| \mu - \frac{1}{N} \sum_{i=1}^N U_i \right\|^2 \leq \mathbb{E}_{V_1,\ldots,V_N} \left\| \mu - \frac{1}{N} \sum_{i=1}^N V_i \right\|^2.$$

Proof. Let $(V_1, \ldots, V_N)$ be i.i.d. as stated. Then

$$\begin{aligned}
\mathbb{E}_{V_1,\ldots,V_N} \left\| \mu - \frac{1}{N}\sum_{i=1}^N V_i \right\|^2 &= \mathbb{E}_{V_1,\ldots,V_N} \left\| \frac{1}{N}\sum_{i=1}^N (V_i - \mu) \right\|^2 \\
&= \mathbb{E}_{V_1,\ldots,V_N} \frac{1}{N^2} \left(\sum_{i=1}^N \|V_i - \mu\|^2 + \sum_{i\neq j} \langle V_i - \mu, V_j - \mu \rangle \right) \\
&= \mathbb{E}_V \frac{1}{N} \|V - \mu\|^2 \\
&= \mathbb{E}_V \frac{1}{N} \left(\|V\|^2 - \|\mu\|^2 \right) \\
&\leq \mathbb{E}_V \frac{1}{N} \|V\|^2 \leq \frac{\sup_{U\in S} \|U\|^2}{N}.
\end{aligned}$$

To conclude, there must exists $(U_1, \ldots, U_N)$ in S so that $\left\| \mu - \frac{1}{N}\sum_{i=1}^N U_i \right\|^2 \leq \mathbb{E}_{V_1,\ldots,V_N} \left\| \mu - \frac{1}{N}\sum_{i=1}^N V_i \right\|^2$. ("Probabilistic method") □

Proposition 3. *Given $U \subset \mathbb{R}^n$,*

$$\mathbb{E}\left(\sup_{a\in U} \left(\frac{1}{n}\sum_{i=1}^n \xi_i a_i \right) \right) \leq \inf_{\epsilon>0} \left(\frac{\epsilon}{\sqrt{n}} + \frac{1}{n} \left(\sup_{a\in U} \|a\|_2 \right) \sqrt{2\log \mathcal{N}(\epsilon, U, \|\cdot\|_2)} \right).$$

Proof. Let $\epsilon > 0$ be arbitrary, and suppose $\mathcal{N}(\epsilon, U, \|\cdot\|_2) < \infty$. Let V be the minimal cover, and $V(a)$ be its closest element to $a \in U$. Then

$$\begin{aligned}
\mathbb{E} \sup_{a\in U} \left(\frac{1}{n}\sum_{i=1}^n \xi_i a_i \right) &= \frac{1}{n} \mathbb{E} \sup_{a\in U} \langle \xi, a - V(a) + V(a) \rangle \\
&= \frac{1}{n} \mathbb{E} \sup_{a\in U} \left(\langle \xi, a \rangle + \|\epsilon\| \, \|a - V(a)\| \right) \\
&\leq \frac{1}{n} \mathbb{E} \sup_{a\in V} \left(\sum_{i=1}^n \xi_i a_i \right) + \frac{\epsilon}{\sqrt{n}} \\
&\leq \frac{1}{n} \left(\sup_{a\in V} \|a\|_2 \right) \sqrt{2\log |V|} + \frac{\epsilon}{\sqrt{n}} \\
&\leq \frac{1}{n} \left(\sup_{a\in U} \|a\|_2 \right) \sqrt{2\log |V|} + \frac{\epsilon}{\sqrt{n}}
\end{aligned}$$

□

4 Generalization error for multi layer network, Radamecher complexity

We will give one Rademacher complexity bound, obtained by inductively peeling off layers. This will depend on $\left\|W_i^\top\right\|_{1,\infty}$. This bound has a pretty clean proof, and appeared in [2].

4.1 First "layer peeling" proof: $(1, \infty)$ norm

Theorem 4. *Let L-Lipschitz activations σ_i satisfy $\sigma_i(0) = 0$, and*

$$\mathcal{F} = \left\{ f : f(x) = \sigma_K(W_K \sigma_{K-1}(\cdots \sigma_1(W_1 x) \cdots)), \ \left\|W_k^\top\right\|_{1,\infty} \leq B \right\},$$

i.e., $\mathcal{F} = \mathcal{F}_\sigma^{(K)}$ with the norm $\|W_k\| = \left\|W_k^\top\right\|_{1,\infty}$. Then

$$\mathsf{Rad}(\mathcal{F}, Z) \leq \|Z\|_{2,\infty} (2LB)^K \sqrt{2\log d}.$$

Remark 5. Many newer bounds replace $\left\|W_k^\top\right\|$ with a distance to initialization. (The NTK is one regime where this helps.)

$(LB)^K$ is roughly a Lipschitz constant of the network according to ∞-norm bounded inputs. Ideally we'd have "average Lipschitz" not "worst case", but we're still far from that.

The factor 2^K is not good and the next section removes it.

We'll prove this with an induction "peeling" off layers. This peeling uses Lemma 1.2, which collects many standard Rademacher properties.

Proof. Let $\mathcal{F}_k$ be the collection of functions computed by each node in kth layer, i.e.,

$$\begin{aligned}\mathcal{F}_k &= \left\{(f_\theta^{(k)})_j : \mathbb{R}^{m^{(k-1)}} \to \mathbb{R} : f_\theta^{(k)} \in \mathcal{F}_\sigma^{(k)}, j = 1, \ldots, m^{(k)}\right\} \\ &= \left\{x \mapsto \sigma_k(W_k f_\theta^{(k-1)}(x)),\ f_\theta^{(k-1)} \in \mathcal{F}_\sigma^{(k-1)}, \left\|W_k^\top\right\|_{1,\infty} \le B\right\}.\end{aligned}$$

It'll be shown by induction that

$$\mathsf{Rad}(\mathcal{F}_k, Z) \le \frac{\|Z\|_{2,\infty} (2LB)^k \sqrt{2\log d}}{n}.$$

Base case ($k = 0$):

$$\begin{aligned}\mathsf{Rad}(\mathcal{F}_0, Z) &= \mathsf{Rad}\left(\{x \mapsto x_j, 1 \le j \le d\}, Z\right) \\ &= \mathbb{E}\left[\sup_{1\le j\le d} \frac{1}{n}\sum_{i=1}^n \xi_i Z_{ij} \,\middle|\, Z^n\right] \\ &\le \frac{1}{n}\left(\max_{1\le j\le d} \left\|Z^j\right\|_2\right)\sqrt{2\log d} \\ &= \frac{\|Z\|_{2,\infty}\sqrt{2\log d}}{n}.\end{aligned}$$

Inductive step. Note that

$$\mathcal{F}_k = \left\{x \mapsto \left(\sigma_k(W_k f_\theta^{(k-1)}(x))\right)_j,\ f_\theta^{(k-1)} \in \mathcal{F}_\sigma^{(k-1)}, \|W_k\| \le B\right\}.$$

Now for $a \in \mathbb{R}^{m_{(k-1)}}$ with $\|a\|_1 \le 1$,

$$a^\top f_\theta^{(k-1)}(x) = \sum_{j=1}^{m_{(k-1)}} a_j (f_\theta^{(k-1)})_j(x),$$

so

$$x \mapsto a^\top f_\theta^{(k-1)}(x) \in \mathrm{conv}(-\mathcal{F}_k \cup \mathcal{F}_k).$$

Now note that, for W_k with $\left\|W_k^\top\right\|_{1,\infty} \le B$, $\frac{W_k}{B}$ has its each row with 1 norm bounded by 1, so $x \mapsto \frac{1}{B} W_k f_\theta^{(k-1)}(x)$ has its component functions (i.e., $\left(\frac{1}{B} W_k f_\theta^{(k-1)}(x)\right)_j$) in $\mathrm{conv}(-\mathcal{F}_k \cup \mathcal{F}_k)$. And therefore,

$$\mathcal{F}_k = \left\{x \mapsto \sigma(Bg(x)),\ g \in \mathrm{conv}(-\mathcal{F}_{k-1} \cup \mathcal{F}_{k-1})\right\}.$$

Hence by applying Lipschitz peeling (Lemma 1.2 (d)),

$$\begin{aligned}\mathsf{Rad}(\mathcal{F}_k, Z) &= \mathsf{Rad}\left(\{x \mapsto \sigma(Bg(x)),\ g \in \mathrm{conv}(-\mathcal{F}_{k-1} \cup \mathcal{F}_{k-1})\}, Z\right) \\ &\le LB\mathsf{Rad}\left(-\mathcal{F}_{k-1} \cup \mathcal{F}_{k-1}, Z\right).\end{aligned}$$

And then from $0 \in \mathcal{F}_k$, from multi-part lemma (Lemma 1.2 (e)),

$$\begin{aligned}\mathsf{Rad}(\mathcal{F}_k, Z) &\le LB\mathsf{Rad}\left(-\mathcal{F}_{k-1} \cup \mathcal{F}_{k-1}, Z\right). \\ &\le (2LB)\mathsf{Rad}\left(\mathcal{F}_{k-1}, Z\right) \\ &\le \frac{(2LB)^k \|Z\|_{2,\infty}\sqrt{2\log d}}{n}.\end{aligned}$$

□

4.2 Second "layer peeling" proof: Frobenius norm

Theorem 6 ([3]). *Let 1-Lipschitz homogeneous activation σ_k be given, and*

$$\mathcal{F} = \left\{ f : f(x) = \sigma_K(W_K \sigma_{K-1}(\cdots \sigma_1(W_1 x) \cdots)),\ \left\|W_k^\top\right\|_F \le B \right\},$$

i.e., $\mathcal{F} = \mathcal{F}_\sigma^{(K)}$ with the norm $\|W_k\| = \|W_k\|_F$. Then

$$\mathsf{Rad}(\mathcal{F}, Z) \le \frac{B^K \|Z\|_F \left(1 + \sqrt{2K \log 2}\right)}{n}.$$

Remark 7. The criticisms of the previous layer peeling proof still apply, except we've removed 2^K.

5 Generalization error for multi layer network, covering number

We will give two generalization bounds.

- The first will be for arbitrary Lipschitz functions, and will be horifically loose (exponential in dimension).
- The second will be, afaik, the tightest known bound for ReLU networks.

5.1 First covering number bound: Lipschitz functions

This bound is intended as a point of contrast with our deep network generalization bounds.

Theorem 8. *Let $R, B > 0$, and let $\mathcal{F}$ denote all L-lipschitz functions from $[-R, +R]^d \to [B, B]$, where Lipschitz is measured with respect to $\|\cdot\|_\infty$. Then the covering number $\mathcal{N}$ satisfies*

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le \max\left\{ 0, \left\lceil \frac{4L(R+\epsilon)}{\epsilon} \right\rceil^d \log \left\lceil \frac{2B}{\epsilon} \right\rceil \right\}.$$

Remark 9. Exponential in dimension.

Revisiting the "point of contrast" comment above, our deep network generalization bounds are polynomial and not exponential in dimension; consequently, we really are doing much better than simply treating the networks as arbitrary Lipschitz functions.

5.2 "Spectrally-normalized" covering number bound

Theorem 10 ([1]). *Fix multivariate activations $\{\sigma_k\}_{k=1}^K$ with σ_k being L_k-Lipschitz and $\sigma_k(0) = 0$, and fix data $Z \in \mathbb{R}^{n \times d}$, and define*

$$\mathcal{F}_n \coloneqq \left\{ \sigma_K(W_K \sigma_{K-1}(\cdots \sigma_1(W_1 Z^\top) \cdots)),\ \left\|W_k^\top\right\|_2 \le s_i, \left\|W_k^\top\right\|_{2,1} \le b_i \right\},$$

and all matrix dimensions are at most m. Then

$$\log \mathcal{N}\left(\epsilon, \mathcal{F}_n, \|\cdot\|_F\right) \le \frac{\|Z\|_F^2 \prod_{k=1}^K L_k^2 s_k^2}{\epsilon^2} \left(\sum_{k=1}^K \left(\frac{b_k}{s_k} \right)^{2/3} \right)^3 \log(2m^2).$$

Remark 11. Applying Proposition 3 gives

$$\mathsf{Rad}(\mathcal{F}, Z) = \tilde{O}\left(\frac{\|Z\|_F}{n} \prod_{k=1}^K L_k s_k \left(\sum_{k=1}^K \left(\frac{b_k}{s_k} \right)^{2/3} \right)^{3/2} \right).$$

Proof uses $\|\sigma(M) - \sigma(M')\|_F \le L \|M - M'\|_F$ if σ is L-Lipschitz; in particular, it allows multi-variate gates like max-pooling.

Let's compare to our best "layer peeling" proof from Theorem 6, which had $\prod_k \|W_k\|_F \lesssim m^{K/2} \prod_k \|W_k\|_2$. If we assume $L_j = 1$, then the comparison becomes

$$m^{K/2} \left(\prod_k \|W_k\|_2 \right) \qquad \text{vs.} \qquad \left(\sum_k \frac{\|W_k^\top\|_{2,1}^{2/3}}{\|W_k\|_2^{2/3}} \right)^{3/2} \left(\prod_k \|W_k\|_2 \right).$$

Then from $K \leq \sum_k \frac{\|W_k^\top\|_{2,1}^{2/3}}{\|W_k\|_2^{2/3}} \leq K m^{2/3}$, the bound is better but still leaves a lot to be desired, and is loose in practice.

The proof, as with Rademacher peeling proofs, is an induction on layers, similarly one which does not "coordinate" the behavior of the layers; this is one source of looseness.

The first step of the proof is a covering number for individual layers.

Lemma 12. *Fora fixed $Z \in \mathbb{R}^{n \times d}$,*

$$\log \mathcal{N}\left(\epsilon, \left\{ WZ^\top : \left\|W^\top\right\|_{2,1} \leq b, W \in \mathbb{R}^{m \times d} \right\}, \|\cdot\|_F \right) \leq \left\lceil \frac{\|Z\|_F^2 b^2}{\epsilon^2} \right\rceil \log(2dm).$$

Proof. Let $W \in \mathbb{R}^{m \times d}$ be given with $\left\|W^\top\right\|_{2,1} \leq b$. Define $s_{ij} \coloneqq W_{ij} / |W_{ij}|$, and note that

$$WZ^\top = \sum_{i,j} e_i e_i^\top W e_j e_j^\top Z^\top = \sum_{i,j} e_i W_{ij} (Ze_j)^\top = \sum_{i,j} \frac{|W_{ij}| \|Ze_j\|_2}{b \|Z\|_F} \frac{b \|Z\|_F s_{ij} e_i (Ze_j)^\top}{\|Ze_j\|_2}.$$

Note by Cauchy-Schwarz that

$$\sum_{i,j} q_{ij} \leq \frac{1}{b\|Z\|_F} \sum_i \sqrt{\sum_j W_{ij}^2} \sqrt{\sum_j \|Ze_j\|_2^2} = \frac{\left\|W^\top\right\|_{2,1} \|Z\|_F}{b\|Z\|_F} \leq 1,$$

potentially with strict inequality, which we will want later. To remedy this, construct probability vector p from q by adding in, with equal weight, some U_{ij} and $-U_{ij}$, so that the above summation form of $WZ^\top$ goes through equally with p as with q.

Now define i.i.d. random variables $(V_1, \ldots, V_N)$, where

$$\begin{aligned}
P(V_l = U_{ij}) &= p_{ij}, \\
\mathbb{E} V_l &= \sum_{i,j} p_{ij} U_{ij} = \sum_{i,j} q_{ij} U_{ij} = WZ^\top, \\
\|U_{ij}\| &= \left\| \frac{s_{ij} e_i (Ze_j)^\top}{\|Ze_j\|_2} \right\|_F \cdot b\|Z\|_F = |s_{ij}| \|e_i\|_2 \left\| \frac{Ze_j}{\|Ze_j\|_2} \right\|_2 b\|Z\|_F = b\|Z\|_F, \\
\mathbb{E}\|V_l\|^2 &= \sum_{i,j} p_{ij} \|U_{ij}\|^2 \leq \sum_{i,j} p_{ij} b^2 \|Z\|_F^2 = b^2 \|Z\|_F^2.
\end{aligned}$$

By Lemma 2, there exist $(\hat{V}_1, \ldots, \hat{V}_N)$ with

$$\left\| WZ^\top - \frac{1}{N} \sum_l \hat{V}_l \right\|^2 \leq \mathbb{E} \left\| \mathbb{E} V_1 - \frac{1}{N} \sum_l V_l \right\|^2 \leq \frac{1}{N} \mathbb{E}\|V_1\|^2 \leq \frac{b^2 \|Z\|_F^2}{N}.$$

Furthermore, the matrices $\hat{V}_l$ have the form

$$\frac{1}{N} \sum_l \hat{V}_l = \frac{1}{N} \sum_l \frac{s_l e_{i_l} (Ze_{j_l})^\top}{\|Ze_{j_l}\|_2} = \frac{1}{N} \sum_l \frac{s_l e_{i_l} e_{j_l}^\top}{\|Ze_{j_l}\|_2} Z^\top;$$

by this form, there are at most $(2md)^N$ choices for $\hat{V}_1, \ldots, \hat{V}_N$. □

Lemma 13. *Let $\mathcal{F}_n$ be the same image vectors as in Theorem 10, and let per-layer tolerances $(\epsilon_1, \ldots, \epsilon_K)$ be given. Then*

$$\log \mathcal{N}\left(\sum_{k=1}^{K} L_k \epsilon_k \prod_{j=k+1}^{K} L_k s_k, \mathcal{F}_n, \|\cdot\|_F\right) \leq \sum_{k=1}^{K} \left\lceil \frac{\|Z\|_F^2 \prod_{j<k} L_j^2 s_j^2}{\epsilon_k^2} \right\rceil \log(2m^2).$$

Sketch of the proof. Let $Z^{(k)}$denote the output of layer k of the network, using weights $(W_k, \ldots, W_1)$, meaning

$$Z^{(0)} \coloneqq Z \qquad \text{and} \qquad Z^{(k)} \coloneqq \sigma_k(Z^{(k-1)} W_k^\top).$$

The proof recursively constructs cover elements $\hat{Z}^{(k)}$ and weights $\hat{W}_k$ for each layer with the following basic properties.

- Define $\hat{Z}^{(0)} \coloneqq Z^{(0)}$, and $\hat{Z}^{(k)} \coloneqq \prod_{B_k} \sigma_k(\hat{Z}^{(k-1)} \hat{W}_k^\top)$, where B_k is the Frobenius-norm ball of radius $\|Z\|_F \prod_{j<k} L_j s_j$.
- Due to the projection $\prod_{B_k}$, $\left\|\hat{Z}^{(k)}\right\|_F \leq \|Z\|_F \prod_{j<k} L_j s_j$. Similarly, using $\sigma_k(0) = 0$, $\left\|Z^{(k)}\right\|_F \leq \|Z\|_F \prod_{j<k} L_j s_j$.
- Given $\hat{Z}^{(k-1)}$, choose $\hat{W}_k$ via Lemma 12 so that $\left\|\hat{Z}^{(k-1)} W_k^\top - \hat{Z}^{(k-1)} \hat{W}_k^\top\right\|_F \leq \epsilon_k$, whereby the corresponding covering number $\mathcal{N}_k$ for this layer satisfies

$$\log \mathcal{N}_k \leq \left\lceil \frac{\left\|\hat{Z}^{(k-1)}\right\|_F^2 b_k^2}{\epsilon_k^2} \right\rceil \log(2m^2) \leq \left\lceil \frac{\|Z\|_F^2 b_k^2 \prod_{j<k} L_j^2 s_j^2}{\epsilon_k^2} \right\rceil \log(2m^2).$$

- Since each cover element $\hat{Z}_k$ depends on the full tuple $(\hat{W}_k, \ldots, \hat{W}_1)$, the final cover is the product of the individual covers (and not their union), and the final cover log cardinality is upper bounded by

$$\log \prod_{k=1}^{K} \mathcal{N}_k \leq \sum_{k=1}^{K} \left\lceil \frac{\|Z\|_F^2 b_k^2 \prod_{j<k} L_j^2 s_j^2}{\epsilon_k^2} \right\rceil \log(2m^2).$$

It remains to prove, by induction, an error guarantee

$$\left\|Z^{(k)} - \hat{Z}^{(k)}\right\|_F \leq \sum_{j=1}^{k} L_j \epsilon_j \prod_{i=j+1}^{k} L_i s_i.$$

which is omitted here. □

Once these Lemmas are shown, the proof for Theorem 10 is a parameter optimization, and I will omit the proof.

References

[1] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *CoRR*, abs/1706.08498, 2017.

[2] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[3] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Inf. Inference*, 9(2):473–504, 2020.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning—with applications in R.* Springer Texts in Statistics. Springer, New York, [2021] ©2021. Second edition [of 3100153].