CSIE5002 Prediction, Learning, and Games: Final Project Report 1–11

# Review of Generalization Bounds for Neural Networks by Rademacher Complexity

**Chien-Min Yu** R07922080@NTU.EDU.TW
*r07922080*

**Ching-Yuan Bai** B05502055@NTU.EDU.TW
*b05502055*

**Wei-I Lin** B05902042@NTU.EDU.TW
*b05902042*

## Abstract

We review several papers that study the generalization bounds for neural networks. The bounds introduced here are upper bounds on empirical Rademacher complexity, meaning that they are data-dependent. We show that these bounds are width-independent, and can be further improved to be depth-independent.

## 1. Introduction

Recently, neural networks have become the focus of study by many due to the broad success in applications. Overparameterized neural networks are able to fit any labels (Zhang et al. (2017)), which gives the natural question to their generalization property. To address this issue, a great number of works have been proposed. One way to bound the generalization error is by giving bounds on a scale-sensitive version of VC-dimension, called the fat-shattering dimension (Bartlett (1997); Anthony and Bartlett (1999)). These analyses can be adapted to bounding the Rademacher complexity (Bartlett and Mendelson (2002); Neyshabur et al. (2015); Bartlett et al. (2017)). Another line of research uses PAC-Bayes analysis to provide generalization guarantees, which builds upon perturbation bounds (Langford and Caruana (2002); Dziugaite and Roy (2017); Neyshabur et al. (2018); Zhou et al. (2019)).

This paper focuses on generalization bounds by Rademacher complexity, and is organized as follows. Section 2 defines some notations used throughout the paper and reviews standard generalization error bounds. Section 3 reviews two bounds on Rademacher complexity and their proof techniques. Section 4 is devoted to depth-independent bounds introduced in Golowich et al. (2018). Finally, some closing remarks are given in Section 5.

## 2. Preliminaries

### 2.1. Vectors and Matrices

For a vector $x \in \mathbb{R}^n$, $\|x\|_p$ denotes the $\ell_p$ norm. For a matrix $X \in \mathbb{R}^{n \times m}$,

- $\|X\|_p$, where $p \in [1, \infty)$, denotes the Schatten-$p$ norm. Specifically, we use $\|X\|$ to denote the spectral norm ($p = \infty$), and $\|X\|_F$ to denote the Frobenious norm ($p = 2$).

- $\|X\|_{p,q} \coloneqq \|(\|X_{:,1}\|_p, \ldots, \|X_{:,m}\|_p)\|_q$ denotes the $(p, q)$ matrix norm.

### 2.2. Neural Networks

Consider the domain $\mathcal{X} = \{\|x\| \leq B\}$ in Euclidean space, a neural network is the function of the following form

$$x \mapsto W_d \sigma_{d-1}(\ldots \sigma_1(W_1 x) \ldots))$$

where each $W_j$ is a parameter matrix and each $\sigma_j$ is some Lipschitz function with $\sigma_j(0) = 0$. Here, $d$ denotes the depth of the network. To simplify the notation, let $W_a^b$ be the matrix tuple $\{W_a, \ldots, W_b\}$, and $N_{W_a^b}$ be the subnetwork between $a$-th layer to $b$-th layer, i.e., the function

$$x \mapsto W_b \sigma_{b-1}(\ldots \sigma_a(W_a x) \ldots)).$$

The function class of neural networks with depth $d$, width $H$, output dimension $K$, and activation function $\sigma_1 = \cdots = \sigma_{d-1} = \sigma$ is defined as

$$N^{d,H,K,\sigma} \coloneqq \{N_{W_1^d} : W_1 \in \mathbb{R}^{H \times D}, \ldots, W_{d-1} \in \mathbb{R}^{H \times H}, W_d \in \mathbb{R}^{K \times H}\}$$

### 2.3. Generalization Error Bound by Rademacher Complexity

Given data $\mathbb{Z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ sampled i.i.d. from an unknown distribution, consider a hypothesis class $\mathcal{H}$ and a loss function $\ell : \mathcal{H} \times \mathbb{Z} \to \mathbb{R}$. For any hypothesis $h \in \mathcal{H}$, the risk is defined as $R(h) \coloneqq \mathbb{E}_z[l(h, z)]$, and the empirical risk $\hat{R}(h) \coloneqq \frac{1}{m}\sum_{i=1}^m \ell(h, z_i)$. Define the function class $\mathcal{F} \coloneqq \{f : z \mapsto \ell(h, z) \mid h \in \mathcal{H}\}$, its Rademacher complexity $\mathcal{R}_m(\mathcal{F}) \coloneqq \mathbb{E}_{\boldsymbol{z},\boldsymbol{\sigma}}[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)]$, and its empirical Rademacher complexity $\hat{\mathcal{R}}_m(\mathcal{F}) \coloneqq \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)]$. The excess risk (or generalization error) $R(h) - \hat{R}(h)$ can be bounded by the Rademacher complexity using standard techniques as in Mohri et al. (2012). Specifically, suppose that the loss function $\ell$ takes values in $[0, 1]$, and we have the following lemmas:

**Lemma 1** *With probability at least $1 - \delta$, every $h \in \mathcal{H}$ satisfies*

$$R(h) \leq \hat{R}(h) + 2\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m}}.$$

**Lemma 2** *With probability at least $1 - \delta$, we have*

$$\mathcal{R}_m(\mathcal{F}) \leq \hat{\mathcal{R}}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m}}.$$

**Lemma 3** *With probability at least $1 - \delta$, every $h \in \mathcal{H}$ satisfies*

$$R(h) \leq \hat{R}(h) + 2\hat{\mathcal{R}}_m(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Lemma 1 and 2 can be proved by McDiarmid's inequality and symmetrization. Combining these two lemmas gives Lemma 3.

## 3. Width-independent Bounds

The goal now is to bound the empirical Rademacher complexity of the neural network class. In this section, we introduce two bounds provided by Neyshabur et al. (2015) and Bartlett et al. (2017), which doesn't have dependence on the width $H$ outside of log terms. They do, however, depend on the depth $d$, whether explicitly or implicitly, even after all log terms are dropped.

### 3.1. Depth-dependent Bound

In this subsection, we deal with binary classification problems ($K = 1$). In order to remove the dependence on width, the network is required to subject to norm constraints. Consider the group-norm regularizer, parametrized by $1 \leq p, q < \infty$

$$\mu_{p,q}(W) = \left( \sum_{k=1}^{d} \sum_{i=1}^{H} \left( \sum_{j=1}^{H} |W_k[i,j]|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}$$

where $W$ is the weight of neural network $f_W$.

**Lemma 4** *Define* $\gamma_{p,q}(W) = \prod_{k=1}^{d} \|W_k\|_{p,q}$. *For all* $f_W \in N^{d,H,K,RELU}$ *and* $1 \leq p, q, \leq \infty$,

$$\gamma_{p,q}(W) = \left( \frac{\mu_{p,q}(W)}{d^{1/q}} \right)^d$$

The proof for Lemma 4 relies on the homogeneity of the ReLU activation function allows the network to rescale the norm of each layer while maintaining identical output. By Lemma 4, we can now consider regularizers defined in terms of $\gamma_{p,q}$ for convenience when using ReLU activation, albeit $\mu_{p,q}$ being the more commonly seen version.

**Lemma 5** *For any* $1 \leq p, q \leq \infty$ *and any set* $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^D$,

$$\hat{\mathcal{R}}_m(N^{1,H,1,RELU}_{\gamma_{p,q} \leq \gamma}) \leq \gamma \max_i \|x_i\|_{p^*} \cdot \max \left\{ \sqrt{\frac{\min\{p^*, 4\log(2D)\}}{m}}, \frac{\sqrt{2}}{m^{1/p}} \right\}$$

*where* $\frac{1}{p^*} + \frac{1}{p} = 1$.

**Proof** (Lemma 5) (sketch)
Note that for any $f_w \in N^1$, $\gamma_{p,q} = \|w\|_p$. ($N^1$ is a set of linear functions)
For a set $S = \{x_1, \ldots, x_m\}$,

$$\begin{aligned}
\hat{\mathcal{R}}_m(N^1_{\gamma_{p,q} \leq \gamma}) &= \mathbb{E}_{\xi \in \{\pm 1\}^m} \left[ \frac{1}{m} \sup_{\|w\|_p \leq \gamma} \left| \sum_{i=1}^{m} \xi_i w^\top x_i \right| \right] \\
&= \mathbb{E}_{\xi \in \{\pm 1\}^m} \left[ \frac{1}{m} \sup_{\|w\|_p \leq \gamma} \left| w^\top \sum_{i=1}^{m} \xi_i x_i \right| \right] \\
&= \gamma \mathbb{E}_{\xi \in \{\pm 1\}^m} \left[ \frac{1}{m} \left\| \sum_{i=1}^{m} \xi_i x_i \right\|_{p^*} \right]
\end{aligned}$$

For $1 \leq p \leq \min\{2, \frac{2\log(2D)}{2\log(2D)-1}\}$, by Massart's Lemma,

$$\mathbb{E}_{\xi\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \xi_i x_i\right\|_{p^*}\right] \leq \max_i \|x_i\|_{p^*}\sqrt{\frac{2D}{m}}$$

For $\min\{2, \frac{2\log(2D)}{2\log(2D)-1}\} < p < \infty$, by Khintchine-Kahane Inequality, Minskowski inequality and subadditivity of polynomial functions,

$$\mathbb{E}_{\xi\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \xi_i x_i\right\|_{p^*}\right] \leq \sqrt{2}\max_i \|x_i\|_{p^*}\frac{1}{m^{1/p}}$$

■

**Theorem 6** *For any $d, p, q \geq 1$ and any set $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^D$*

$$\hat{\mathcal{R}}_m(N^{d,H,1,RELU}_{\gamma_{p,q}\leq\gamma}) \leq \gamma(2H^{[\frac{1}{p^*}-\frac{1}{q}]_+})^{d-1}\cdot D_0$$

*where $D_0 = \max\{\sqrt{\frac{\min\{p^*, 2\log(2D)\}\cdot\sup\|x_i\|_{p^*}^2}{m}}, \frac{\sqrt{2}\gamma\sup\|x_i\|_{p^*}}{m^{1/p}}\}$ and $\frac{1}{p}+\frac{1}{p^*}=1$.*

**Proof** (Theorem 6) (sketch)
Here we prove by induction on depth $d$. Induction basis holds by Lemma 5. For the induction step, by the Contraction Lemma of Lipschitz functions,

$$\hat{\mathcal{R}}_m(N^{d,H,1,\text{RELU}}_{\gamma_{p,q}\leq\gamma}) \leq 2H^{[\frac{1}{p^*}-\frac{1}{q}]_+}\hat{\mathcal{R}}_m(N^{d-1,H,1,\text{RELU}}_{\gamma_{p,q}\leq\gamma})$$

■

**Corollary 7** *For any $d, p \geq 1$, $1 \leq q \leq p^* = \frac{p}{p-1}$ and any set $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^D$*

$$\hat{\mathcal{R}}_m(N^{d,H,1,\sigma_{RELU}}_{\gamma_{p,q}\leq\gamma}) \leq \gamma 2^{d-1}\cdot D_0$$

*where $D_0 = \max\{\sqrt{\frac{\min\{p^*, 2\log(2D)\}\cdot\sup\|x_i\|_{p^*}^2}{m}}, \frac{\sqrt{2}\gamma\sup\|x_i\|_{p^*}}{m^{1/p}}\}$*

Directly applying Theorem 6 with such $q$ yields Corollary 7.

We close this subsection by making several remarks of Corollary 7.

- **Small** norm is necessary to properly bound the empirical Rademacher complexity of infinite-width network with weight norms as intuitively summation or product of large norms is unable to bound the width factor.
- The proof structure of induction on depth while explicitly bounding the empirical Rademacher complexity of each layer produces some factor per layer. Thus, another proof framework is required to remove exponential dependence on the bound (for example, frameworks that bound the empirical Rademacher complexity of the network as a whole).

### 3.2. Almost Depth-independent Bound

In this subsection we deal with a general multiclass classification problem. The loss function under consideration is the *ramp loss* $\ell_\gamma : \mathbb{R} \to \mathbb{R}^+$ defined as

$$\ell_\gamma(r) \coloneqq \begin{cases} 0 & r < -\gamma, \\ 1 + r/\gamma & r \in [-\gamma, 0], \\ 1 & r > 0. \end{cases}$$

The *ramp risk* is defined as $R_\gamma(f_W) \coloneqq \mathbb{E}\left[\ell_\gamma\left(\max_{i \neq y} f_W(x)_i - f_W(x)_y\right)\right]$, and the empirical counterpart $\hat{\mathcal{R}}_\gamma(f_W)$ is defined similarly by replacing the expectation with the finite sum. Notice that $\mathcal{R}_\gamma$ is an upper bound for the probability of error on the source distribution, which can be combined with Lemma 3 to give the following corollary:

**Corollary 8** *Given $\gamma > 0$, define*

$$\mathcal{F}_\gamma \coloneqq \left\{ (x, y) \mapsto \ell_\gamma \left( \max_{i \neq y} f_W(x)_i - f_W(x)_y \right) : f_W \in N^{d,H,K,\sigma} \right\}.$$

*Then with probability at least $1 - \delta$, every network $f_W \in N^{d,H,K,\sigma}$ satisfies*

$$\Pr\left[ \operatorname*{argmax}_i f_W(x)_i \neq y \right] \leq \hat{R}_\gamma(f_W) + 2\hat{\mathcal{R}}_m(\mathcal{F}_\gamma) + 3\sqrt{\frac{\log(2/\delta)}{2m}}.$$

The reason to bound the excess ramp risk (instead of the excess $0/1$ risk) is due to its Lipschitz continuity, a property that will come in handy. The next step is to bound the empirical Rademacher complexity through Dudley entropy integral. Let the covering number $\mathcal{N}(U, \epsilon, \|\cdot\|)$ be defined as

$$\mathcal{N}(U, \epsilon, \|\cdot\|) \coloneqq \inf_{V \subseteq U} \left\{ |V| : \sup_{A \in U} \inf_{B \in V} \|A - B\| \leq \epsilon \right\},$$

and we have the following lemma:

**Lemma 9** *Suppose that $\mathcal{F}$ consists of functions taking values in $[0, 1]$ and that $\mathbf{0} \in \mathcal{F}$. Let $\mathcal{F}_m$ denote the image of some given data of size $m$ under $\mathcal{F}$. We have*

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

Lemma 9 can be proved by standard techniques (chaining), even though the data metric is not normalized by $\sqrt{m}$. One main contribution of Bartlett et al. (2017) is to provide an upper bound on this covering number. We now state the theorem formally and provide a proof sketch.

**Theorem 10** *Assume the activation function $\sigma$ is 1-Lipschitz and $\sigma(\mathbf{0}) = \mathbf{0}$. Let data matrix $X \in \mathbb{R}^{m \times D}$, spectral norm bounds $(s_1, \ldots, s_d)$ and matrix $(2,1)$ norm bounds $(b_1, \ldots, b_d)$ be given. Then for any $\epsilon > 0$,*

$$\ln \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \leq \frac{\|X\|_2^2 \ln(2H^2)}{\epsilon^2} \left( \prod_{j=1}^d s_j^2 \right) \left( \sum_{j=1}^d \left( \frac{b_j}{s_j} \right)^{2/3} \right)^3,$$

*where*

$$\mathcal{H}_X \coloneqq \left\{ f_W(X^T) : f_W \in N^{d,H,K,\sigma}, \|W_j\| \leq s_j, \|W_j^T\|_{2,1} \leq b_j \right\}.$$

**Proof** (Theorem 10) (sketch) First, a matrix covering number bound can be proved by Maurey's sparsification lemma. Then, a covering number bound for the entire network can be proved by induction on layers. ■

An example of the activation function satisfying the constraints is the RELU function. The above choice of $(2,1)$ norm is crucial to prevent combinatorial parameters ($H$ and $d$) from appearing outside of log terms. We are now ready to provide a generalization error bound, which requires the data bound and norm constraints to be given beforehand.

**Corollary 11** *Assume the activation function $\sigma$ is 1-Lipschitz and $\sigma(\mathbf{0}) = \mathbf{0}$. Let margin $\gamma$, data bound $B$, spectral norm bounds $(s_1, \ldots, s_d)$ and matrix $(2,1)$ norm bounds $(b_1, \ldots, b_d)$ be given. Then with probability at least $1-\delta$, every network $f_W \in N^{d,H,K,\sigma}$ with $\|W_j\| \le s_j, \|W_j^T\|_{2,1} \le b_j$ satisfies*

$$\Pr\left[\operatorname*{argmax}_i f_W(x)_i \neq y\right] \le \hat{R}_\gamma(f_W) + \frac{8}{m} + \frac{72B\ln(2W)\ln(m)}{\gamma m}\left(\prod_{j=1}^d s_j\right)\left(\sum_{j=1}^d \left(\frac{b_j}{s_j}\right)^{2/3}\right)^{3/2} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

**Proof** (Corollary 11) Notice that $x \mapsto \ell_\gamma\left(\max_{i\neq y} f_W(x)_i - f_W(x)_y\right)$ is $2/\gamma$-Lipschitz wrt $\|\cdot\|_p$ for any choice of $y$, $f_W$, and $p$. Therefore,

$$\ln \mathcal{N}\left((\mathcal{F}_\gamma)_m, \epsilon, \|\cdot\|_2\right) \le \ln \mathcal{N}\left(\mathcal{H}_X, \frac{\gamma}{2}\epsilon, \|\cdot\|_2\right).$$

It remains to apply Theorem 10, Lemma 9, and Corollary 8. ■

Notice that the bound in the right-hand side of Corollary 11 is indeed increasing in each $s_j$, since

$$\left(\prod_{j=1}^d s_j\right)\left(\sum_{j=1}^d \left(\frac{b_j}{s_j}\right)^{2/3}\right)^{3/2} = \left(\sum_{j=1}^d \left(b_j \prod_{i\neq j} s_i\right)^{2/3}\right)^{3/2}.$$

Dividing the parameter space of Corollary 11 and applying the union bound, we get the following theorem. See Bartlett et al. (2017) for the complete proof.

**Theorem 12** *Assume the activation function $\sigma$ is 1-Lipschitz and $\sigma(\mathbf{0}) = \mathbf{0}$. With probability at least $1-\delta$, every margin $\gamma > 0$ and network $f_W \in N^{d,H,K,\sigma}$ satisfy*

$$Pr\left[\operatorname*{argmax}_i f_W(x)_i \neq y\right] \le \hat{R}_\gamma(f_W) + \tilde{\mathcal{O}}\left(\frac{\|X\|_2 R_W}{\gamma m} + \sqrt{\frac{\ln(1/\delta)}{m}}\right),$$

*where*

$$R_W := \left(\prod_{j=1}^d \|W_j\|\right)\left(\sum_{j=1}^d \left(\frac{\|W_j^T\|_{2,1}}{\|W_j\|}\right)^{2/3}\right)^{3/2}.$$

We close this subsection by making several remarks of Theorem 12.

- This bound doesn't have any combinatorial factors outside of log terms, including the width $H$, the depth $d$, and the output dimension $K$. Therefore, this improves upon the bound given in Theorem 6.
- For the regression setting ($K = 1$), we can consider the $\ell_2$ loss (Brier loss). By assuming that each $y$ and $f_W(x)$ lies in the range $[-C, C]$, the loss function takes values in $[0, 4C^2]$ and is $(2C)$-Lipschitz continuous. Following the same derivation, we can provide a similar bound for the excess $\ell_2$ risk.

## 4. From Depth-dependent to Depth-independent Bounds

Yet we derive an width-independent Rademacher complexity bound in Section 3, the bound in Theorem 12 still depends on the network depth implicitly: To observe this, note that $\|W_j\|_{2,1} \geq \|W_j\|_F \geq \|W_j\|$ holds for any $W_j$, so the bounds cannot be smaller than

$$\tilde{\mathcal{O}}\left(B\left(\prod_{j=1}^{d}\|W_j\|\right)\sqrt{\frac{d^3}{m}}\right).$$

In this section, we introduce a general technique proposed by Golowich et al. (2018), that allows us to convert a depth-dependent Rademacher complexity bound to a depth-independent one, assuming some control over any Schatten-$p$ norm of the parameter matrices.

We provide a formal description of the technique and a sketch of its proof in Section 4.1. In Section 4.2, we provide an application to the technique, obtaining depth-independent bounds on the sample complexity of various classes of neural networks. Please refer to Golowich et al. (2018) for the proofs of the theorems in this section if they are not provided.

### 4.1. General Result

To obtain the general result, the procedure is as follows: Let a class of depth-$d$ networks and some $r \in \{1, \ldots, d\}$ be given.

1. (Theorem 13) Relate the class of depth-$d$ networks to the class of similar networks with specific properties.
2. (Theorem 15) With the specific properties, bound the sample complexity of the class of similar networks with that of a class of depth-$r'$ networks, for some $1 \leq r' \leq r$.
3. (Theorem 16) Derive the sample complexity bound of the class of depth-$d$ networks that relates only to the parameter $r$.

As $r$ is arbitrary, optimizing over $r$ yields a bound independent to $d$.

We begin with the ideas that for any network there exists a similar network with a specific layer condition, as formalized in the following theorem and lemma.

**Theorem 13** *For any $p \in [1, \infty[$, any network $N_{W_1^d}$ such that $\prod_{j=1}^{d}\|W_j\| \geq \Gamma$ and $\prod_{j=1}^{d}\|W_j\|_p \leq M$, and for any $r \in \{1, \ldots, d\}$, there exists another network $N_{\bar{W}_1^d}$ with the following properties:*

- *$\tilde{W}_1^d$ is identical to $W_1^d$, except for the parameter matrix $\tilde{W}_{r'}$ in the $r'$-th layer, for some $r' \in \{1, 2, \ldots, r\}$. The matrix $\tilde{W}_{r'}$ is of rank at most 1, and equals $suv^T$ where $s, u, v$ are some leading singular value and singular vectors pairs of $W_{r'}$.*
- $\sup_{x \in \mathcal{X}} \|N_{W_1^d}(x) - N_{\tilde{W}_1^d}(x)\| \leq B \left( \prod_{j=1}^d \|W_j\| \right) \left( \frac{2p \log M/\Gamma}{r} \right)^{1/p}$.

**Lemma 14** *Let $\ell$ be a $\gamma$-Lipschitz loss function from Euclidean space to $\mathbb{R}$ and $\mathcal{H}, \mathcal{H}'$ be two hypothesis class from $\mathcal{X}$ to Euclidean space such that for any $h \in \mathcal{H}$ and $h' \in \mathcal{H}'$, $\sup_{x \in \mathcal{X}} \|h(x) - h'(x)\| \leq B$ holds for some $B \geq 0$, then*

$$\hat{\mathcal{R}}_m(\ell \circ \mathcal{H}) \leq \hat{\mathcal{R}}_m(\ell \circ \mathcal{H}') + \gamma B.$$

With Theorem 13, we observe that the network $N_{\tilde{W}_1^d}$:

$$x \mapsto W_d \sigma_{d-1}(\ldots \sigma_{r'}(suv^T \sigma_{r'-1}(\ldots \sigma_1(W_1 x) \ldots)))$$

can be decomposed into the composition of the depth-$r'$ network

$$x \mapsto v^T \sigma_{r'-1}(\ldots \sigma_1(W_1 x) \ldots)),$$

and the univariate function

$$x \mapsto W_d \sigma_{d-1}(\ldots \sigma_{r'}(sux)).$$

Moreover, due to the constraints on the product of the parameter matrices' Schatten-$p$ norms and the Lipschitz constants of the activation functions, the latter function is Lipschitz.

By Lemma 14, it suffices to consider the Rademacher complexities of a subset of the class of depth-$r'$ networks composed with univariate Lipschitz functions. In fact, given any class of bounded functions, one can bound the Rademacher complexity of its composition with univariate Lipschitz functions, as formalized in the following theorem.

**Theorem 15** *Let $\mathcal{H}$ be a class of functions from Euclidean space to $[-R, R]$. Let $\mathcal{F}_{L,a}$ be the class of of $L$-Lipschitz functions from $[-R, R]$ to $\mathbb{R}$, such that $f(0) = a$ for some fixed $a$. Let $\mathcal{F}_{L,a} \circ \mathcal{H} := \{f(h(\cdot)) : f \in \mathcal{F}_{L,a}, h \in \mathcal{H}\}$, then its Rademacher complexity satisfies:*

$$\hat{\mathcal{R}}_m(\mathcal{F}_{L,a} \circ \mathcal{H}) \leq cL \left( \frac{R}{\sqrt{m}} + \log^{3/2}(m) \hat{\mathcal{R}}_m(\mathcal{H}) \right),$$

*where $c$ is a universal constant.*

Combining the ideas above, we have the following theorem, which allows us to bound the sample complexity of a class of depth-$d$ network with no dependence on $d$. It is the main result in Golowich et al. (2018).

**Theorem 16** *Consider the following hypothesis class of networks on $\mathcal{X} = \{x : \|x\| \leq B\}$:*

$$\mathcal{H} = \left\{ N_{W_1^d} : \begin{array}{c} \prod_{j=1}^d \|W_j\| \geq \Gamma, \\ \forall j \in \{1, \ldots, d\}\ W_j \in \mathcal{W}_j, \|W_j\| \leq M(j), \|W_j\|_p \leq M_p(j) \end{array} \right\}$$

*for some parameters* $p, \Gamma \geq 1$ *and* $\{M(j), M_p(j), \mathcal{W}_j\}_{j=1}^d$. *Also, for any* $r \in \{1, \dots d\}$, *define*

$$\mathcal{H}_r = \left\{ N_{W_1^r} : \begin{array}{c} N_{W_1^r} \text{ maps to } \mathbb{R},\ \prod_{j=1}^r \|W_j\| \geq \Gamma, \\ \forall j \in \{1, \dots, r-1\}\ W_j \in \mathcal{W}_j, \\ \forall j \in \{1, \dots, r\}\ \|W_j\| \leq M(j), \|W_j\|_p \leq M_p(j) \end{array} \right\}.$$

*For any* $m > 1$, *let* $l \circ \mathcal{H} = \{l(h(\cdot)) : h \in \mathcal{H}\}$, *where* $l$ *is a* $\frac{1}{\gamma}$*-Lipschitz real-valued functions with* $l(0) = a$, *for some* $a \in \mathbb{R}$ *satisfying* $|a| \leq \frac{B}{\gamma} \prod_{j=1}^d M(j)$. *Then* $\hat{\mathcal{R}}_m(l \circ \mathcal{H})$ *is upper bounded by*

$$\frac{cB}{\gamma} \prod_{j=1}^d M(j) \left( \frac{\log^{3/2}(m)}{B} \max_{r' \in \{1,\dots,r\}} \frac{\hat{\mathcal{R}}_m(\mathcal{H}_{r'})}{\prod_{j=1}^{r'} M(j)} + \left( \frac{\log\left(\frac{1}{\Gamma} M_p(j)\right)}{r} \right)^{1/p} + \frac{1 + \sqrt{\log r}}{\sqrt{m}} \right)$$

*for any* $r \in \{1, \dots, d\}$, *where* $c > 0$ *is a universal constant.*

### 4.2. Depth-Independent Bound

In this section, we apply Theorem 16 to Corollary 11, which provides empirical Rademacher complexity of a depth-$d$ network with some norm constraints, obtaining a depth-independent bound on the empirical Rademacher complexity of the neural network.

To derive the bound, we will use the following trick.

**Lemma 17** *For any* $\alpha > 0, \beta \in ]0, 1]$ *and* $b, c, n \geq 1$, *it holds that*

$$\min_{r \in \{1,\dots,d\}} \frac{cr^\alpha}{n} + \frac{b}{r^\beta} \leq 3 \cdot \frac{b^{\frac{\alpha}{\alpha+\beta}}}{(n/c)^{\frac{\beta}{\alpha+\beta}}}.$$

Now, plugging in Corollary 11 to bound $\hat{\mathcal{R}}_m(\mathcal{H}_{r'})$ in Theorem 16 gives a bound dependent on $r$, where $r \in \{1, \dots d\}$. Then, by optimizing over $r$ with Lemma 17, we obtain the following corollary.

**Corollary 18** *Let* $\mathcal{H}$ *be the class of depth-*$d$*, width-*$h$ *networks with* 1*-Lipschitz, positive-homogeneous, element-wise activation functions. Assuming the loss function* $\ell$ *is a* $\frac{1}{\gamma}$*-Lipschitz real-valued functions with* $\ell(0) = a$, *for some* $a \in \mathbb{R}$ *satisfying* $|a| \leq \frac{B}{\gamma} \prod_{j=1}^d M(j)$, *and the hypothesis class* $\mathcal{H}$ *defined as*

$$\mathcal{H} = \left\{ N_{W_1^d} : \begin{array}{c} \prod_{j=1}^d \|W_j\| \geq \Gamma, \\ \forall j \in \{1, \dots, d\}\ \|W_j^T\|_{2,1} \leq M_{2,1}(j), \|W_j\| \leq M(j), \|W_j\|_p \leq M_p(j) \end{array} \right\},$$

*it holds that the empirical Rademacher complexity* $\hat{\mathcal{R}}_m(\ell \circ \mathcal{H})$ *is at most*

$$\mathcal{O}\left( \frac{BL \log(h) \log(m) \prod_{j=1}^d M(j)}{\gamma} \cdot \frac{\bar{\log}\left(\frac{1}{\Gamma} \prod_{j=1}^d M_p(j)\right)^{\frac{1}{\frac{2}{3}+p}} \left(\bar{\log}^{3/2}(m)\right)^{\frac{1}{1+\frac{3}{2}p}}}{m^{\frac{1}{2+3p}}} \right)$$

*where* $\bar{\log}(z) \coloneqq \max\{1, \log(z)\}$.

Ignoring the logarithmic factors, the above bound becomes

$$\tilde{\mathcal{O}}\left(\frac{BL\prod_{j=1}^{d}M(j)}{\gamma}\sqrt[4]{\frac{\left(\frac{1}{\Gamma}\prod_{j=1}^{d}M_p(j)\right)}{\sqrt{m}}}\right),$$

which is independent on $d$.

**Remark 19** *If we take $p = 2$, then Corollary 18 is almost a depth-independent bound with the same assumptions in Corollary 11 proposed by Bartlett (1997). Here the term "almost" cannot be "exactly" because the former requires the activation function to be positive-homogeneous, which is stronger than $\ell(0) = 0$, assumed in the latter.*

## 5. Concluding Remarks

To this end, we introduce two depth-dependent bounds and shows how to convert them into depth-independent bounds. The results seem satisfying in terms of big-Oh; however, when computed on real data sets, these bounds might still be very loose. To get generalization bounds that are non-vacuous, Zhou et al. (2019) combine the PAC-Bayes framework and compression approach (Arora et al. (2018)), and demonstrate its tightness on ImageNet classification problem.

To explain the generalization ability of overparametrized networks, another fruitful research direction is to take the learning algorithm into consideration (Allen-Zhu et al. (2018); Allen-Zhu and Li (2019)). Indeed, hoping that the norm doesn't scale with the width $H$ is essentially requiring that the weight matrices to be sparse, which reduces to the non-overparameterized case more or less. It is interesting to investigate whether considering both the learning algorithm (e.g. SGD) and the neural network class may still give generalization bounds that are independent of the number of neurons.

## References

Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD learn recurrent neural networks with provable generalization? *CoRR*, abs/1902.01028, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *CoRR*, abs/1811.04918, 2018.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, pages 254–263, 2018.

Peter L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9*, pages 134–140. MIT Press, 1997.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299, 2018.

John Langford and Rich Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, pages 809–816, 2002.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, pages 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*, 2019.