Notes on compression based bounds for noncompressed networks

Simon Vary

November 29, 2019

These notes are concerned with generalization error for noncompressed neural networks which can be well compressed as shown in [10].

1 Introduction

The classical learning theory suggests that overparameterized models cause overfitting. However, deep neural networks are empirically observed to generalize well despite having far more parameters than the number of training samples.

A common explanation of this phenomenon is that there is a regularization during the training, either implicit by the use of SGD or explicit by weight decay, dropout, or batch normalization. This argument is questioned by the experimental evidence [12] that standard neural nets trained with SGD and explicit regularization can still achieve low training error on randomly labeled examples that don't generalize. Clearly, deep nets trained on real-life data have some properties not achieved by only imposing regularization, which reduce effective capacity of the model.

Using a compression based bounds is a promising approach for tight generalization error evaluation. These bounds measure how much the trained neural network can be compressed and characterize the the size of the neural network as its effective dimension [1]. Compression based bounds guarantee the generalization error of the compressed neural network, and fail to give an explanation for why large networks can avoid overfitting.

The presented paper *Compression based bound for noncompressed network: unified generalization error analysis of a large compressible deep neural network* [10] obtains a compression based bound for a noncompressed network (but which can be well compressed) and thus it gives an explanation why deep learning generalizes well despite its large network size. However, the paper does not answer the question why gradient descent training on real data tends to converge to a network that is well compressible (either by sparse pruning or low-rank thresholding on weight matrices).

The main result of the paper is the following bound

$$\Psi(\widehat{f}) \leq \hat{\Psi}(\hat{f}) + \underbrace{2\bar{R}_n(\hat{\mathcal{G}}) + \sqrt{2tM}\frac{1}{\sqrt{n}}}_{\text{main term}} + C\underbrace{\left[\dot{R}_n(\hat{\mathcal{F}} - \widehat{\mathcal{G}})\log(n)^{\frac{3}{2}} + \dot{r}\sqrt{t}\frac{1}{\sqrt{n}} + (1+tM)\frac{1}{n}\right]}_{\text{fast term}}, \tag{1}$$

with probability at least $1 - 3e^{-t}$ for all $t \geq 1$, where $\Psi(\widehat{f})$ is the generalization error of the noncompressed estimator $\hat{f} \in \hat{\mathcal{F}}$, and $\hat{\mathcal{G}}$ is the class of compressed estimators.

The result will be discussed in detail, but roughly if $\dot{r}^2 \gtrsim \|\hat{f} - \hat{g}\|_n$ is of order $\mathcal{O}_p(1)$ (i.e. noncompressed $\hat{f}$ is not far from compressed $\hat{g}$), and $\dot{R}_n(\hat{\mathcal{F}} - \widehat{\mathcal{G}})$ is of order $\mathcal{O}(1/\sqrt{n})$, then the fast term can be faster than the main term which is $\mathcal{O}(1/\sqrt{n})$ ([10] argues this is true in a typical compressed net setting). If one neglect these terms, the bound can be written as

$$\Psi(\widehat{f}) \leq \hat{\Psi}(\widehat{f}) + O_p\left(\bar{R}_n(\widehat{\mathcal{G}}) + \frac{1}{\sqrt{n}}\|\widehat{f} - \widehat{g}\|_n + \frac{1}{\sqrt{n}}\right), \tag{2}$$

which is asymptotically faster than the compression based bound on generalisation error of the compressed network $\hat{g}$:

$$\Psi(\widehat{g}) \leq \widehat{\Psi}(\widehat{f}) + \|\widehat{f} - \widehat{g}\|_n + C\bar{R}_n(\widehat{\mathcal{G}}), \tag{3}$$

by $1/\sqrt{n}$ of the bias term (i.e. the compression error) $\|\hat{f} - \hat{g}\|_n$.

1.1 Preliminaries

Consider the standard supervised learning problem formulation where each data point consists of an input $x \in \mathbb{R}^d$ and a label $y \in \mathbb{R}$. We are given n, i.i.d. observations $D_n = (x_i, y_i)_{i=1}^n \sim P$ distributed from a probability distribution P. The marginal distributions are denoted as $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$.

To measure the performance of a trained function f, such that $f(x) \approx y$, we use a loss function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and define the training error and the expected error as

$$\widehat{\Psi}(f) := \frac{1}{n}\sum_{i=1}^{n} \psi\left(y_i, f\left(x_i\right)\right), \tag{4}$$

$$\Psi(f) := \mathrm{E}_{Z\sim P}[\psi(Y, f(X))]. \tag{5}$$

The main task is to bound the generalization error $\Psi(\hat{f}) - \widehat{\Psi}(\hat{f})$ for an estimator $\hat{f}$. Define the following norms

$$\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2}, \qquad \text{emprical } L_2\text{-norm given } D_n \tag{6}$$

$$\|f\|_{L_2} := \sqrt{\mathrm{E}_{Z\sim P}[f(Z)^2]}, \qquad \text{population } L_2\text{-norm.} \tag{7}$$

We deal with deep neural networks as a model. The activation function is denoted as ν and is 1−Lipschitz. The depth of the network is $L \in \mathbb{N}$ and the width of the ℓ-th layer is m_ℓ ($\ell = 1, \ldots, L+1$) where we set $m_1 = d$ (dimension of input) and $m_{L+1} = 1$ (dimension of output). Denote the set of networks with depth L and widths $\mathbf{m} = (m_1, \ldots, m_L)$ with norm constraints as

$$\begin{aligned} \mathrm{NN}\left(\mathbf{m}, R_2', R_{\mathrm{F}}'\right) := \{ f(x) = G \circ (W^{(L)}\eta(\cdot)) \circ (W^{(L-1)}\eta(\cdot)) \circ \cdots \circ (W^{(1)}x) \mid \\ W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell+1}}, \|W^{(\ell)}\|_2 \leq R_2', \|W^{(\ell)}\|_{\mathrm{F}} \leq R_{\mathrm{F}}' \Big\}, \end{aligned} \tag{8}$$

where $\|W\|_2$ is the operator norm, $\|W\|_F$ is the Frobenius norm, and G is the "clipping" operator $G(x) = \max\{-M, \min\{x, M\}\}$ for a constant M.

We denote $\mathcal{F} = \mathrm{NN}(\mathbf{m}, R_2, R_F)$ to represent the full model for a given $R_2, R_F > 0$. Throughout the analysis it is assumed that $R_2 \approx 1$ but R_F may be moderately large.

Rademacher complexity of a function class $\mathcal{F}'$ is denoted as

$$\hat{R}_n(\mathcal{F}') := \mathrm{E}_\varepsilon \left[\sup_{f \in \mathcal{F}'} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(z_i) | D_n \right], \qquad \text{conditional Rademacher complexity} \tag{9}$$

$$\bar{R}_n(\mathcal{F}') := \mathrm{E}_{D_n}\left[\hat{R}_n(\mathcal{F}') \right], \qquad \text{Rademacher complexity} \tag{10}$$

The difficulty lies in that $\bar{R}(\mathcal{F})$ for the function class $\mathrm{NN}(\mathbf{m}, R_2', R_F')$ is too large and the naive VC bound on the error is loose. An approach to this issue is a compression based bound.

Here is a brief summary of the generalization error bounds for noncompressed networks. **Naive VC bound** [3] gives $\bar{R}_n(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{L\sum_{\ell=1}^n m_\ell m_{\ell+1}}{n}\log(n)}\right)$, which is impractical because the number of parameters in the numerators is usually much larger then n. **Norm-based bound** [5] gives $\bar{R}_n(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{L(R_F)^L}{n}}\right)$, which grows exponentially with the depth L resulting in a loose bound. The work of [9] showed a norm based bound which eliminated the exponantial dependency on L and gives $\bar{R}_n(\mathcal{F}) = \mathcal{O}\left(\sqrt{\frac{L^3(\max_\ell m_\ell)R_{\mathrm{F}}^2/R_2^2}{n}}\right)$, however there is dependency on the width $\sum_\ell m_\ell (R_F)^2$ which is larger than linear order of the width since R_F can be moderately large. Other works derive that bounds can avoid the exponential dependence on the depth L but will depend linearly or quadratically on the maximal width even though R_2 is bounded. For a more complete overview look in Section 2 of [10].

1.2 Compression based bound

Suppose that the trained network $\hat{f}$ is in the subset of the neural network model: $\hat{f} \in \hat{\mathcal{F}} \subset \mathcal{F}$. For example, $\mathcal{F}$ can be a set of networks with weight matrices that are approximately rank restricted and have bounded norms.

The goal is to give a uniform bound, valid for any $\hat{f} \in \hat{\mathcal{F}}$ that satisfies the condition that it can be well approximated by a compressed network $\hat{g}$ which is included in a more restricted compressed submodel $\hat{g} \in \hat{\mathcal{G}}$.

The compression based bounds are given as

$$\Psi(\hat{g}) \leq \widehat{\Psi}(\hat{f}) + \underbrace{\|\hat{f} - \hat{g}\|_n}_{\text{compression error}} + C\bar{R}_n(\hat{\mathcal{G}}), \tag{11}$$

for a constant $C > 0$. The term $\|\hat{f} - \hat{g}\|_n$ adapts the empirical error of $\hat{f}$ to that of $\hat{g}$. For an overview of generalization error bounds for compressed networks look at Table 1 in [10], or in Appendix B.

Although, the right-hand side of (11) depends on the complexity of $\hat{\mathcal{G}}$ which is assumed to be much smaller than that of the full model $\mathcal{F}$, the left hand side is not the expected error of $\hat{f}$ but that of $\hat{g}$. One way to transfer $\Psi(\hat{f})$ is to use Lipschitz continuity of the loss function $|\Psi(\hat{g}) - \Psi\hat{f}| \leq \|\hat{g} - \hat{f}\|_{L_2}$ to convert the bound in (11) to

$$\Psi(\widehat{f}) \leq \widehat{\Psi}(\widehat{f}) + \left(\|\widehat{f} - \widehat{g}\|_n + \|\widehat{f} - \widehat{g}\|_{L_2}\right) + \bar{R}_n(\widehat{\mathcal{G}}). \tag{12}$$

However, to bound the term $\|\widehat{f} - \widehat{g}\|_n + \|\widehat{f} - \widehat{g}\|_{L_2}$, there typically appears the complexity of the model $\mathcal{F}$ which is larger than the compressed model $\mathcal{G}$

$$\|\hat{f} - \hat{g}\|_n \leq \sqrt{\|\hat{f} - \hat{g}\|_{L_2}^2 + \mathcal{O}_p(\bar{R}(\hat{\mathcal{F}}))}, \tag{13}$$

and results in a slow convergence rate.

This can be overcome by carefully controlling the difference between the training and test error $\hat{f}$ and $\hat{g}$ by utilizing the local Rademacher complexity

$$\dot{R}_r\left(\mathcal{F}'\right) := \bar{R}_n\left(\left\{f \in \mathcal{F}' | \|f\|_{L_2} \leq r\right\}\right). \tag{14}$$

Assumption 1 (Lipschitz continuity). The loss function ψ is 1-Lipschitz continuous with respect to the function output:

$$|\psi(y, u) - \psi(y, u')| \leq |u - u'| \quad (\forall y \in \text{supp}(P_{\mathcal{Y}}), u, u' \in \mathbb{R}). \tag{15}$$

The activation function ν is also 1-Lipschitz continuous: $\|\nu(u) - \nu(u')\| \leq \|u - u'\|, (\forall u \in \mathbb{R}^{d'})$ where d' is any positive integer.

Assumption 2 (Bounded input). The norm of input is bounded by $B_x > 0 : \|x\| \leq B_x$, $(\forall x \in \text{supp}(P_{\mathcal{X}}))$.

Assumption 3 (Bounded norms of $\mathcal{F}$ and $\mathcal{G}$). The L_∞-norms of all elements in $\hat{\mathcal{F}}$ and $\hat{\mathcal{G}}$ are bounded by $M \geq 1 : \|f\|_\infty, \|g\|_\infty \leq M$ for all $f \in \hat{\mathcal{F}}$ and $g \in \hat{\mathcal{G}}$.

2 Compression bound for noncompressed network

Denote the Minkowski difference of estimator classes $\mathcal{F}$ and $\mathcal{G}$ by $\mathcal{F}-\mathcal{G} := \left\{ f-g \,|\, f \in \hat{\mathcal{F}}, g \in \hat{\mathcal{G}} \right\}$. Assume that the local Rademacher complexity of the set $\mathcal{F}-\mathcal{G}$ is concave in terms of $r>0$: Supposed that there exists $\phi : [0,\infty) \to [0,\infty)$ such that

$$\dot{R}_r(\widehat{\mathcal{F}}-\widehat{\mathcal{G}}) \le \phi(r) \qquad \text{and} \qquad \phi(2r) \le 2\phi(r) \quad (\forall r>0). \tag{16}$$

Define $r_* = r_*(t)$ as

$$r_*(t) := \inf \left\{ r>0 \,\middle|\, 8\frac{\phi(r)}{r^2} + M\sqrt{\frac{4t}{r^2 n}} + M^2 \frac{2t}{r^2 n} \le \frac{1}{2} \right\} \tag{17}$$

Theorem 1 (Compression bound for noncompressed network)**.** *Suppose that the empricical L_2-distance between $\hat{f}$ and $\hat{g}$ is bounded by $\|\hat{f}-\hat{g}\|_n \le \hat{r}^2$ for a fixed $\hat{r}>0$ almost surely. Let $\dot{r} := \sqrt{2(\hat{r}^2 + r_*^2)}$, then, under Assumptions 1, 2, 3, there exists a universal constant $C>0$ such that*

$$\Psi(\widehat{f}) \le \hat{\Psi}(\hat{f}) + \underbrace{2\bar{R}_n(\hat{\mathcal{G}}) + \sqrt{2tM}\frac{1}{\sqrt{n}}}_{\textit{main term}} + C \underbrace{\left[\dot{R}_n(\hat{\mathcal{F}}-\widehat{\mathcal{G}}) \log(n)^{\frac{3}{2}} + \dot{r}\sqrt{t}\frac{1}{\sqrt{n}} + (1+tM)\frac{1}{n} \right]}_{\textit{fast term}}, \tag{18}$$

with probability at least $1-3e^{-t}$ for all $t \ge 1$.

Proof. Will be presented as in Appendix A of [10]. □

Furthermore, the bound can be refined by directly evaluating the covering number of $\hat{\mathcal{F}}-\widehat{\mathcal{G}}$. What follows is one such refinement for a special case when the weight matrices are approximately of low-rank.

Assumption 4 Assume that each of weight matrices $W^{(\ell)}$, $(\ell = 1, \ldots, L)$ of any $f \in \hat{\mathcal{F}}$ is near low rank, that is, there is exists $\alpha > 1/2$ and $V_0 > 0$ such that

$$\sigma_j\left(W^{(\ell)}\right) \le V_0 j^{-\alpha}, \tag{19}$$

where $\sigma_j(W)$ is the j^{th} largest singular values of a matrix W (so $\sigma_1 \ge \sigma_2(W) \ge \ldots \ge 0$).

Theorem 2. *The compressed model $\hat{\mathcal{G}} = \mathrm{NN}(\boldsymbol{m}, s, R_2, R_F)$ has the following complexity:*

$$\bar{R}_n(\widehat{\mathcal{G}}) \le CM\sqrt{\frac{\sum_{\ell=1}^{L} s_\ell \left(m_\ell + m_{\ell+1}\right)}{n} \log(n)}. \tag{20}$$

If $\hat{\mathcal{F}}$ satisfied Assumption 4, we can set

$$\hat{r} = V_0 R_2^{L-1} B_x \sum_{\ell=1}^{L} s_\ell^{-\alpha}, \tag{21}$$

for any $\hat{f} \in \hat{\mathcal{F}}$, there exists $\hat{g} \in \hat{\mathcal{G}}$ such that $\|\hat{f} - \hat{g}\|_n \leq \hat{r}$. Then letting

$$A_1 = L\frac{\sum_{\ell=1}^{L} s_\ell \left(m_\ell + m_{\ell+1}\right)}{n} \log(n) \quad \textit{and} \quad A_2 = L\frac{\left(\sum_{\ell=1}^{L} m_\ell\right)\left(2LV_0 R_2^{L-1} B_x\right)^{1/\alpha}}{n}, \tag{22}$$

the overall generalization error is bounded by

$$\Psi(\widehat{f}) \leq \widehat{\Psi}(\widehat{f}) + C\left[MA_1 + M^{\frac{2\alpha-1}{2\alpha+1}} A_2^{\frac{2\alpha}{1+2\alpha}} + \sqrt{\hat{r}^{2(1-2\alpha)} A_2} + (\hat{r}+M)\sqrt{A_1} + \frac{1+tM}{n}\right], \tag{23}$$

with probability $1 - 3e^{-t}$ for any $t > 1$ where $C > 0$ is a constant depending on α.

If one balances the rank optimally, we get

$$\Psi(\widehat{f}) \leq \hat{\Psi}(\widehat{f}) + C\left[M^{1-1/2\alpha}\sqrt{L\frac{\left(\sum_{\ell=1}^{L} m_\ell\right)\left(2LV_0 R_2^{L-1} B_x\right)^{1/\alpha}}{n}\log(n)} + M^{\frac{2\alpha-1}{2\alpha+1}} A_2^{\frac{2\alpha}{2\alpha+1}} + \frac{1+tM}{n}\right]. \tag{24}$$

The bound in (24) is $\mathcal{O}(\sqrt{L\frac{\sum_{\ell=1}^{L} m_\ell}{n}})$ so has linear dependency on the width m_ℓ. So the compressible model achieves much better generalization than the naive VC-bound which has dependency $\mathcal{O}\left(\sqrt{L\frac{\sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}}\right)$.

References

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, *Stronger generalization bounds for deep nets via a compression approach*, (2018), pp. 1–39.

[2] P. Bartlett, D. J. Foster, and M. Telgarsky, *Spectrally-normalized margin bounds for neural networks*, (2017), pp. 1–24.

[3] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, *Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks*, (2017), pp. 1–16.

[4] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Non-Asymptotic Theory of Indepence*, Oxford University Press, feb 2012.

[5] N. Golowich, A. Rakhlin, and O. Shamir, *Size-Independent Sample Complexity of Neural Networks*, 75 (2017), pp. 1–3.

[6] S. Ingo and A. Christmann, *Support Vector Machines*, Information Science and Statistics, Springer New York, New York, NY, 2008.

[7] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, 1991.

[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, SSRN Electronic Journal, (2018).

[9] B. Neyshabur, S. Bhojanapalli, and N. Srebro, *A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks*, (2017), pp. 1–9.

[10] T. Suzuki, *Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network*, (2019).

[11] M. Talagrand, *New concentration inequalities in product spaces*, Inventiones Mathematicae, 126 (1996), pp. 505–563.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, (2017), pp. 1–15.

A Lemmata

Lemma A.1 (Rademacher concentration inequality)**.** *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{X}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. samples D_n of size n, the following holds for all $g \in \mathcal{G}$*

$$\mathbb{E}[g(z)] \leq \frac{1}{n}\sum_{i=1}^{n} g\left(z_i\right) + 2\bar{R}_n(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{25}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{n}\sum_{i=1}^{n} g\left(z_i\right) + 2\widehat{R}_n(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{26}$$

Proof. Proof can be found in [8] (Theorem 3.3, page 31). □

Lemma A.2 (Contraction inequality)**.** *Let $x_1, \ldots, x_n$ be vectors whose real-valued components are indexed by $\mathcal{T}$, that is, $x_i = (x_{i,s})_{s\in\mathcal{T}}$. For each $i = 1, \ldots, n$ let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function such that $\phi_i(0) = 0$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, and let $\Psi : [0,\infty) \to \mathbb{R}$ be a non-decreasing convex function. Then*

$$\boldsymbol{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(x_{i,s}\right)\right)\right] \leq \boldsymbol{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i x_{i,s}\right)\right] \tag{27}$$

and

$$\boldsymbol{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(x_{i,s}\right)\right|\right)\right] \leq \boldsymbol{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i x_{i,s}\right|\right)\right] \tag{28}$$

Proof. Proof can be found in [4] (Theorem 11.6, page 315). □

Lemma A.3 (Dudley integral)**.** *Let $\mathcal{F}$ be a real-valued function class taking values in $[0,1]$ and assume that $0 \in \mathcal{F}$. Then*

$$\hat{R}_n(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{\sqrt{n}}\sqrt{\log\mathcal{N}\left(\mathcal{F}_{|D_n}, \varepsilon, \|\cdot\|\right)}d\varepsilon\right). \tag{29}$$

Proof. Proof can be found in appendices of [2]. □

Lemma A.4 (Sudakov's minoration). *Let T be a subset of $\mathbb{R}^n$; for every $\varepsilon > 0$*

$$\varepsilon\sqrt{\log\left(\mathcal{N}\left(T, \|\cdot\|, \varepsilon\right)\right)} \leq K\hat{R}(T)\left(\log\left(2 + \frac{\sqrt{n}}{\hat{R}(T)}\right)\right)^{1/2}, \tag{30}$$

where $K > 0$ is some numerical constant.

Proof. Proof can be found in [7] (Corollary 4.14, page 116). □

Lemma A.5 (Talgrand's concentration inequality). *Let $\mathcal{G}$ be a function class on $\mathcal{X}$ that is separable with respect to ∞-norm, and $\{x_i\}_{i=1}^n$ be i.i.d. random variables with values $\mathcal{X}$. Furthermore, let $B \geq 0$ and $U \geq 0$ be $B := \sup_{g\in\mathcal{G}} \mathrm{E}\left[(g - \mathrm{E}[g])^2\right]$ and $U := \sup_{g\in\mathcal{G}}\|g\|_\infty$, then for $Z := \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i) - \mathrm{E}[g]\right|$, we have*

$$P\left(Z \geq 2\mathrm{E}[Z] + \sqrt{\frac{2Bt}{n}} + \frac{2Ut}{n}\right) \leq e^{-t}, \tag{31}$$

for all $t > 0$.

Proof. Proof can be found in [11] (Corollary 4.14, page 116). □

Lemma A.6 (Peeling device). *Let $(Z, \mathcal{A}, \mathrm{P})$ be a probability space, (T, d) be a separable metric space, $h : T \to [0, \infty)$ be a continuous function, and $(g_t)_{t\in T} \subset \mathcal{L}_0(Z)$ be Caratheodory family. We define $r^* := \inf\{h(t) : t \in T\}$. Moreover, let $\varphi : (r^*, \infty) \to [0, \infty)$ be a function such that $\varphi(4r) \leq 2\varphi(r)$ and*

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{\substack{t\in T \\ h(t)\leq r}} |g_t(z)| \leq \varphi(r), \tag{32}$$

for all $r > r^*$. Then, for all $r > r^*$, we have

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t) + r} \leq \frac{4\varphi(r)}{r}. \tag{33}$$

Proof. Proof can be found in [6] (Theorem 7.7, page 248 and subsequent derivation on page 249). □

B Comparison of generalization error bounds for compressed networks (Table 1 from [10])

Table 1: Comparison of each generalization error to our bound. R_{F} is the Frobenius norm of the weight matrix, R_2 is the operator norm of the weight matrix, $R_{p\to q}$ is the (p,q) matrix norm, L is the depth, m is the maximum of the width, n is the sample size. $\bar{R}_n$ and $\dot{R}_r$ represent the Rademacher complexity and local Rademacher complexity respectively. κ is a Lipschitz constant between layers. α represents the eigenvalue drop rate of the weight matrix, and β represents that of the covariance matrix among the nodes in each internal layer. $\hat{r}$ is the bias induced by compression. "Original" indicates whether the bound is about the original network or not.

Authors	Rate	Bound type	Original
Neyshabur et al. [39]	$\frac{2^L R_{\mathrm{F}}^L}{\sqrt{n}}$	Norm base	Yes
Bartlett et al. [6]	$\frac{R_2^L}{\sqrt{n}}\left(L\frac{R_{2\to1}^{2/3}}{R_2^{2/3}}\right)^{3/2}$	Norm base	Yes
Wei & Ma [56]	$\frac{\left(1+L\kappa^{\frac{4}{3}}R_{2\to1}^{2/3}+L\kappa^{\frac{2}{3}}R_{1\to1}^{2/3}\right)^{3/2}}{\sqrt{n}}$	Norm base	Yes
Neyshabur et al. [40]	$\frac{R_2^L}{\sqrt{n}}\sqrt{L^3 m\frac{R_{\mathrm{F}}^2}{R_2^2}}$	Norm base	Yes
Golowich et al. [18]	$R_{\mathrm{F}}^L \min\left\{\frac{1}{n^{1/4}}, \sqrt{\frac{L}{n}}\right\}$	Norm base	Yes
Li et al. [32] Harvey et al. [21]	$\frac{R_2^L\sqrt{L^2m^2}}{\sqrt{n}}$	VC-dim.	Yes
Arora et al. [1]	$\hat{r}+\sqrt{\frac{L^2\max_{1\le i\le n}\lvert\widehat{f}(x_i)\rvert^2\sum_{\ell=1}^L\frac{1}{\mu_\ell^2\mu_{\ell\to}^2}}{n\hat{r}^2}}$	Compression	No
Suzuki et al. [48]	$\hat{r}+\sqrt{\frac{\sum_{\ell=1}^L m_{\ell+1}^\sharp m_\ell^\sharp}{n}}$	Compression	No
Ours (Thm. 1)	$\hat{r}\sqrt{\frac{1}{n}}+\dot{R}_{\hat{r}}(\widehat{\mathcal{F}}-\widehat{\mathcal{G}})+\bar{R}_n(\widehat{\mathcal{G}})$	General	Yes
Ours (Cor. 1)	$\sqrt{L(L\kappa^2)^{1/\alpha}\frac{Lm}{n}}$	Low rank weight	Yes
Ours (Thm. 4)	$\sqrt{L^{1+\frac{\beta}{\frac{4\alpha}{(2\alpha-1)}+\beta}}\frac{(Lm)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}{n}}$	Low rank weight Low rank cov.	Yes