ROLE OF OVER-PARAMETERIZATION IN GENERALIZATION OF 3-LAYER RELU NETWORKS

Simranjit Singh* simsingh@microsoft.com
Aditya Golatkar* adityagolatkar@ucla.edu
Avijit Verma* avijit2verma@ucla.edu

ABSTRACT

Over-parameterized neural networks defy conventional wisdom by generalizing effectively; however, standard complexity metrics like norms and margins fail to account for this. A recent work introduced a novel measure considering unit-wise capacities and provided a better explanation and tighter generalization bounds but was confined to two-layer networks. This paper extends that framework to three-layer ReLU networks. We empirically confirm the applicability of these measures and introduce a corresponding theoretical Rademacher complexity bound.

1 INTRODUCTION AND RELATED WORK

Deep neural networks have enjoyed great success in a wide variety of tasks. Traditional statistical learning theory suggests that increasing model complexity results in over fitting to the training data. However, for neural networks, it is observed that the generalization error decreases with an increase in the model capacity Neyshabur et al. (2014). Different measures that have been proposed to measure the model capacity like VC dimension, norm, margin and sharpness generally increase with model size and hence fail to explain better generalization of models with increasing size. Neyshabur et al. (2018) proposed novel unit-wise complexity measures, namely, *unit capacity* and *unit impact* of the hidden neurons, which they empirically observed to decrease with increasing model size. However, their work was limited to two-layer ReLU networks.

Motivated by their results, we first empirically investigate the trend of *unit capacity* and *unit impact* of units in both hidden layers of a three-layer ReLU network. We observed that these measures follow trends similar to two layer networks. We also present a Rademacher bound for three layer network and empirically find that this bound decreases with increasing number of hidden units.

2 GENERALIZATION OF THREE-LAYER RELU NETWORKS

Similar to Neyshabur et al. (2018), we consider three layer fully connected neural networks with input dimension d and output dimension c as

$$f_{\mathbf{W},\mathbf{V},\mathbf{U}}(\mathbf{x}) = \mathbf{W}[\mathbf{V}[\mathbf{U}\mathbf{x}]_+]_+$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{U} \in \mathbb{R}^{h_1 \times d}$, $\mathbf{V} \in \mathbb{R}^{h_2 \times h_1}$, $\mathbf{W} \in \mathbb{R}^{c \times h_2}$ and $[]_+$ denotes the element-wise ReLU operation. The task is to classify into c-classes and we use the same ramp loss function as used in Neyshabur et al. (2018) which is defined as:

$$l_\gamma(f(x), y) = \begin{cases} 0 & \mu(f(x), y) > \gamma \\ 1 - \mu(f(x), y)/\gamma & \mu(f(x), y) \in [0, \gamma] \\ 1 & \mu(f(x), y) < 0 \end{cases}$$

where $\mu(f(x), y) = f(x)[y] - \max_{i \neq y} f(x)[i]$. l_γ is a Lipschitz loss function with Lipschitz parameter $\sqrt{2}/\gamma$ and $\gamma = 0$ reduces it to classification loss.

2.1 EMPIRICAL INVESTIGATION

To constrain the Rademacher complexity for three-layer networks, we focus on a narrowed function class suggested by Neyshabur et al. (2018)), emphasizing spectral norm, Frobenius norm, and

*Equal contribution. Work done while authors were at the University of California, Los Angeles (UCLA).

distance Frobenius norm relative to initialization i.e., $\|\mathbf{U} - \mathbf{U_0}\|_F$ for first layer. In Fig 1 we plot these three measures over the weight matrices (Experiment setting details in Appendix). It is clear that the distance Frobenius norm diminishes with more hidden units, hinting at SGD's preference for solutions near the starting parameters. Consequently, our complexity bound should incorporate the distance Frobenius norm of the hidden layers and the Frobenius norm of the output layer.

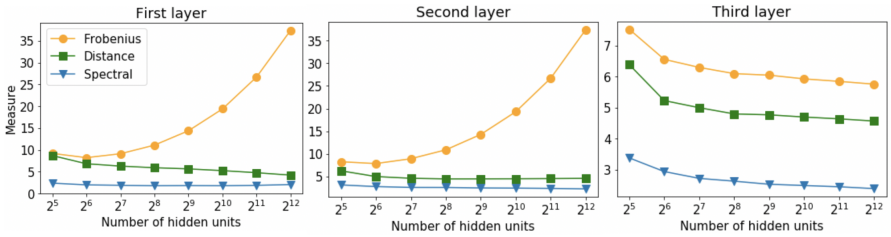

Figure 1: Plot of different measures on a three-layer ReLU network trained on MNIST data set.

2.2 Generalization Bound

We consider a restricted set of parameters below:

$$\Pi = \Big\{(\mathbf{W}, \mathbf{V}, \mathbf{U}) \Big| \mathbf{W} \in \mathbb{R}^{c \times h_2}, \mathbf{V} \in \mathbb{R}^{h_2 \times h_1}, \mathbf{U} \in \mathbb{R}^{h_1 \times d}, \\ \|\mathbf{w}_j\| \leq \delta_j, |v_{j,i} - v^0_{j,i}| \leq \beta_{j,i}, \|\mathbf{u}_i - \mathbf{u}_i{}^0\| \leq \alpha_i \Big\}$$

here, $\mathbf{w}_j$ and $\mathbf{u}_i$ denote the j^{th} column and i^{th} row of $\mathbf{W}$ and $\mathbf{U}$ respectively. Then the function class of interest becomes all the 3-layer ReLU networks with parameters in Π:

$$\mathcal{F}_\Pi = \{f(\mathbf{x}) = \mathbf{W}\big[\mathbf{V}\big[\mathbf{U}\mathbf{x}\big]_+\big]_+ | (\mathbf{W}, \mathbf{V}, \mathbf{U}) \in \Pi\}$$

Then from our observation: $\|\mathbf{U} - \mathbf{U}_0\|_F, \|\mathbf{V} - \mathbf{V}_0\|_F, \|\mathbf{W}\|_F$ decreases with h_1, h_2, while $\|\mathbf{U}\|_F, \|\mathbf{V}\|_F$ increases with h_1, h_2. Now, we first state an important lemma required to bound the Rademacher complexity, based on which we state the theorem below which introduces our Rademacher complexity bound. We provide the proof for both in Appendix.

Lemma 1. *Given a training set $S = \{x_i\}_{i=1}^m$ and $\gamma > 0$, Rademacher complexity of the composition of loss function l_γ over the class $\mathcal{F}_\Pi$ is bounded as follows:*

$$\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{U},\mathbf{V},\mathbf{W}\in\Pi} \frac{6}{\gamma m} \sum_{k=1}^{m} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \Big\langle \sigma_k, v_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k})\right]$$

Theorem 1. *Given a training set $S = \{x_i\}_{i=1}^m$ and $\gamma > 0$, Rademacher complexity of the composition of loss function l_γ over the class $\mathcal{F}_\Pi$ is bounded as follows:*

$$\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \frac{6\sqrt{2}}{\gamma\sqrt{m}} \cdot \sqrt{\sum_{j=1}^{h_2} \delta_j^2} \cdot \sqrt{\sum_{j=1}^{h_2} \sum_{i=1}^{h_1} \beta_{j,i}^2} \cdot \left(\sqrt{\sum_{i=1}^{h_1} \alpha_i^2} \frac{\|\mathbf{X}\|_F}{\sqrt{m}} + \frac{\|\mathbf{U}^0\mathbf{X}\|_F}{\sqrt{m}}\right)$$

The upper bound on Rademacher complexity derived from Theorem 1 decreases with increasing the number of hidden units in the neural network as seen from Figure 2 in Appendix.

3 Conclusion and Future Work

In this work, we derived an upper bound for the Rademacher complexity of three-layer ReLU networks based on distance Frobenius norm of weights measured w.r.t. initialization and showed that this upper bound decreases with an increase in model complexity. This is in line with the practical observation that the generalization bound of the deep neural networks decreases with increase in model complexity for three-layer networks. Our proof was based on the empirical observations we made about spectral, Frobenius and distance norms over different number of hidden units.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *CoRR*, abs/1802.05296, 2018. URL http://arxiv.org/abs/1802.05296.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *CoRR*, abs/1706.08498, 2017. URL http://arxiv.org/abs/1706.08498.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020.

Amit Daniely. SGD learns the conjugate kernel class of the network. *CoRR*, abs/1702.08503, 2017. URL http://arxiv.org/abs/1702.08503.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *CoRR*, abs/1602.05897, 2016. URL http://arxiv.org/abs/1602.05897.

Simon S. Du, Jason D. Lee, Yuandong Tian, Barnabás Póczos, and Aarti Singh. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. *CoRR*, abs/1712.00779, 2017. URL http://arxiv.org/abs/1712.00779.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *CoRR*, abs/1712.06541, 2017. URL http://arxiv.org/abs/1712.06541.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. *CoRR*, abs/1703.02930, 2017. URL http://arxiv.org/abs/1703.02930.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *CoRR*, abs/1806.07572, 2018. URL http://arxiv.org/abs/1806.07572.

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis D. Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *CoRR*, abs/1806.05159, 2018. URL http://arxiv.org/abs/1806.05159.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. *CoRR*, abs/1503.00036, 2015. URL http://arxiv.org/abs/1503.00036.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *CoRR*, abs/1707.09564, 2017. URL http://arxiv.org/abs/1707.09564.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Konstantinos Pitas, Mike E. Davies, and Pierre Vandergheynst. Pac-bayesian margin bounds for convolutional neural networks - technical report. *CoRR*, abs/1801.00171, 2018. URL `http://arxiv.org/abs/1801.00171`.

LeCun Yann, Cortes Corinna, and J Christopher. The mnist database of handwritten digits. *URL http://yhann. lecun. com/exdb/mnist*, 1998.

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, 2019.

A Appendix

A.1 Our Rademacher Complexity bound

In Figure 2, we plot our theoretical bound on Rademacher complexity that we derived in Theorem 1. We see that it does indeed decrease with increase in the number of hidden units in both the hidden layers. This verifies our inferences from our empirical investigation.

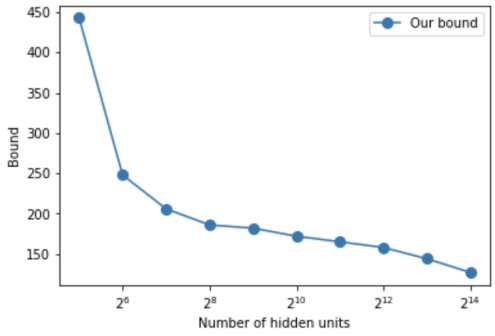

Figure 2: Plot showing that our Rademacher complexity bound decreases with an increase in the number of hidden units.

A.2 Experiments Settings

We trained fully connected three layer ReLU networks on the MNSIT data set Yann et al. (1998). We considered the number of units in the first and the second hidden layers to be equal for simplicity. In all, we trained 13 architectures with sizes from 2^3 to 2^{15} each time increasing the number of hidden units by a factor of 2. We used Stochastic Gradient descent (SGD) with momentum 0.9 and learning rate 0.01 to train the network. The mini batch size was 64. No explicit regularization techniques were used. The stopping criteria was the same as in Neyshabur et al. (2018), i.e., stop the training when the loss reaches 0.01 or when the number of epochs have reached 1000.

We modified the author's original code from Neyshabur et al. (2018) to carry out our experiments. We made changes to the architecture and the measure calculations as suited to our project.

A.3 Related Work

Previous studies have established that the VC-dimension of neural networks scales linearly with the number of parameters at a minimum Harvey et al. (2017), highlighting the inadequacy of classic VC theory to account for the generalization capabilities of contemporary neural networks, which often have more parameters than available training samples. Various researchers have explored norm-based approaches to generalization bounds (as seen in works by Bartlett & Mendelson (2002); Bartlett et al. (2017); Neyshabur et al. (2015),Neyshabur et al. (2017),Neyshabur et al. (2018) ; Pitas et al. (2018); Golowich et al. (2017); Li et al. (2018)), alongside compression-based strategies Arora et al. (2018). Employing the PAC-Bayes methodology, Du et al. (2017) and Zhou et al. (2019) succeeded in deriving non-trivial generalization bounds for networks trained on datasets like MNIST and ImageNet.

A novel analytical perspective has been introduced via the Neural Tangent Kernel (NTK) Jacot et al. (2018), which elucidates the gradient descent dynamics in Artificial Neural Networks (ANNs). This framework bridges ANNs and kernel methods, particularly evident when considering ANNs of infinite width. There have been other methods linking deep learning and kernel methodologies, as explored by Chizat et al. (2020), Daniely et al. (2016) Daniely (2017).

A.4 Proof for Lemma 1

Lemma 1. *Given a training set $S = \{x_i\}_{i=1}^m$ and $\gamma > 0$, Rademacher complexity of the composition of loss function l_γ over the class $\mathcal{F}_\Pi$ is bounded as follows:*

$$\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{U},\mathbf{V},\mathbf{W}\in\Pi} \frac{6}{\gamma m} \sum_{k=1}^{m}\sum_{j=1}^{h_2}\delta_j \sum_{i=1}^{h_1}\Big\langle \sigma_k, v_{j,i}\Big\rangle(\alpha_i\|\mathbf{x}_k\| + \rho_{i,k})\right]$$

We prove this lemma using induction on t for the following result:

Proof. We prove this lemma using induction on t for the following result:

$$m\mathcal{R}_S(\ell_\gamma\circ\mathcal{F}_\Pi) \leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{U},\mathbf{V},\mathbf{W}\in\Pi} \frac{6}{\gamma} \sum_{k=1}^{t-1}\sum_{j=1}^{h_2}\delta_j \sum_{i=1}^{h_1}\Big\langle \sigma_k, v_{j,i}\Big\rangle(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k})+\sum_{k=t}^{m}\sigma_k l_\gamma(\mathbf{W}[\mathbf{V}[\mathbf{U}\mathbf{x}_k]_+]_+, y_k)\right]$$

where $\rho_{i,k} = \left|\left\langle \mathbf{u}_i^0, \mathbf{x}_k\right\rangle\right|$.
We write the above equation more compactly as

$$m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{U},\mathbf{V},\mathbf{W}\in\Pi} \sigma_t l_\gamma(\mathbf{W}[\mathbf{V}[\mathbf{U}\mathbf{x}_t]_+]_+, y_t) + \phi_{\mathbf{U},\mathbf{V},\mathbf{W}}\right]$$

where

$$\phi_{\mathbf{U},\mathbf{V},\mathbf{W}} = \frac{6}{\gamma}\sum_{k=1}^{t-1}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\Big\langle \sigma_k, v_{j,i}\Big\rangle(\alpha_i\|\mathbf{x}_k\| + \rho_{i,k}) + \sum_{k=t+1}^{m}\sigma_k l_\gamma(\mathbf{W}[\mathbf{V}[\mathbf{U}\mathbf{x}_k]_+]_+, y_k)$$

The above statement will hold trivially for the base case when $t = 1$, from the definition of Radamacher complexity.

Let us assume that the above statement is true for any $k < t$ and prove it for $k = t$

$$\begin{aligned} m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) &\leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{U},\mathbf{V},\mathbf{W}\in\Pi} \sigma_t l_\gamma(\mathbf{W}[\mathbf{V}[\mathbf{U}\mathbf{x}_t]_+]_+, y_t) + \phi_{\mathbf{U},\mathbf{V},\mathbf{W}}\right] \\ &\leq \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\Big[\sup_{(\mathbf{W},\mathbf{V},\mathbf{U})\in\Pi}\sup_{(\mathbf{W}',\mathbf{V}',\mathbf{U}')\in\Pi} \ell_\gamma\Big(\mathbf{W}\big[\mathbf{V}\big[\mathbf{U}\mathbf{x}_t\big]_+\big]_+, y_t\Big) - \ell_\gamma\Big(\mathbf{W}'\big[\mathbf{V}'\big[\mathbf{U}'\mathbf{x}_t\big]_+\big]_+, y_t\Big) \\ &\quad + \phi_{\mathbf{U},\mathbf{V},\mathbf{W}} + \phi_{\mathbf{U}',\mathbf{V}',\mathbf{W}'})\Big] \end{aligned}$$

From the Lipschitz property of loss, we get:

$$m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\Big[\sup_{(\mathbf{W},\mathbf{V},\mathbf{U})\in\Pi}\sup_{(\mathbf{W}',\mathbf{V}',\mathbf{U}')\in\Pi} \frac{\sqrt{2}}{\gamma}\underbrace{\Big\|\mathbf{W}\big[\mathbf{V}\big[\mathbf{U}\mathbf{x}_t\big]_+\big]_+ - \mathbf{W}'\big[\mathbf{V}'\big[\mathbf{U}'\mathbf{x}_t\big]_+\big]_+\Big\|_2}_{(A)} + \phi_{\mathbf{U},\mathbf{V},\mathbf{W}} + \phi_{\mathbf{U}',\mathbf{V}',\mathbf{W}'}\Big] \tag{1}$$

To bound (A), we add and subtract $\mathbf{W}'\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_1\right]_+\right]_+$ in (A) and using triangle inequality to obtain:

$$\left\|\mathbf{W}\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+ - \mathbf{W}'\left[\mathbf{V}'\left[\mathbf{U}'\mathbf{x}_t\right]_+\right]_+\right\|_2 \leq \underbrace{\left\|\mathbf{W}\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+ - \mathbf{W}'\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+\right\|_2}_{(B)} + \underbrace{\left\|\mathbf{W}'\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+ - \mathbf{W}'\left[\mathbf{V}'\left[\mathbf{U}'\mathbf{x}_t\right]_+\right]_+\right\|_2}_{(C)} \tag{2}$$

To bound (B), we use the fact that ReLU is 1-Lipschitz along with triangle inequality to obtain:

$$\begin{aligned}\left\|\mathbf{W}\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+ - \mathbf{W}'\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+\right\|_2 &\leq \sum_{j=1}^{h_2}\|\mathbf{w}_j - \mathbf{w}_j^{'}\|\left|\left\langle \mathbf{v}_j, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle\right| \\ &\leq \sum_{j=1}^{h_2}\|\mathbf{w}_j - \mathbf{w}_j^{'}\|\sum_{i=1}^{h_1}|v_{j,i}|\left|\left\langle \mathbf{u}_i, \mathbf{x}_t\right\rangle\right| \\ &\leq \sum_{j=1}^{h_2} 2\delta_j \sum_{i=1}^{h_1}|v_{j,i}|(\alpha_i\|\mathbf{x}_t\| + \rho_{i,t}),\end{aligned} \tag{3}$$

where last step follows from the fact that $\|\mathbf{w}_j\| \leq \delta_j$ and
$\left|\left\langle \mathbf{u}_i, \mathbf{x}_t\right\rangle\right| = \left|\left\langle \mathbf{u}_i - \mathbf{u}_i^0 + \mathbf{u}_i^0, \mathbf{x}_t\right\rangle\right| \leq \left|\left\langle \mathbf{u}_i - \mathbf{u}_i^0, \mathbf{x}_t\right\rangle\right| + \left|\left\langle \mathbf{u}_i^0, \mathbf{x}_t\right\rangle\right| \leq \underbrace{\|\mathbf{u}_i - \mathbf{u}_i^0\|}_{\leq \alpha_i}\|\mathbf{x}_t\| + \rho_{i,t}$

To bound (C), we add and subtract $\left\langle \mathbf{v}_j', \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle$ and use triangle inequality

$$\begin{aligned}&\left\|\mathbf{W}'\left[\mathbf{V}\left[\mathbf{U}\mathbf{x}_t\right]_+\right]_+ - \mathbf{W}'\left[\mathbf{V}'\left[\mathbf{U}'\mathbf{x}_t\right]_+\right]_+\right\|_2 \\ &\leq \sum_{j=1}^{h_2}\|\mathbf{w}_j'\|\left|\left\langle \mathbf{v}_j, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j', \left[\mathbf{U}'\mathbf{x}_t\right]_+\right\rangle\right| \\ &= \sum_{j=1}^{h_2}\|\mathbf{w}_j'\|\left|\left\langle \mathbf{v}_j, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j', \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle + \left\langle \mathbf{v}_j', \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j', \left[\mathbf{U}'\mathbf{x}_t\right]_+\right\rangle\right| \\ &\leq \underbrace{\sum_{j=1}^{h_2}\|\mathbf{w}_j'\|\left|\left\langle \mathbf{v}_j, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j', \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle\right|}_{(D)} + \underbrace{\sum_{j=1}^{h_2}\|\mathbf{w}_j'\|\left|\left\langle \mathbf{v}_j', \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j', \left[\mathbf{U}'\mathbf{x}_t\right]_+\right\rangle\right|}_{(E)}\end{aligned} \tag{4}$$

To bound (D)

$$\begin{aligned}\sum_{j=1}^{h_2}\|\mathbf{w}_j^{'}\|\left|\left\langle \mathbf{v}_j, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle - \left\langle \mathbf{v}_j^{'}, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle\right| &= \sum_{j=1}^{h_2}\|\mathbf{w}_j^{'}\|\left|\left\langle \mathbf{v}_j - \mathbf{v}_j^{'}, \left[\mathbf{U}\mathbf{x}_t\right]_+\right\rangle\right| \\ &\leq \sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}|v_{j,i} - v_{j,i}^{'}|\left|\left\langle \mathbf{u}_i, \mathbf{x}_t\right\rangle\right| \\ &\leq \sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}|v_{j,i} - v_{j,i}^{'}|(\alpha_i\|\mathbf{x}_t\| + \rho_{i,t})\end{aligned} \tag{5}$$

To bound (E)

$$\sum_{j=1}^{h_2} \|\mathbf{w}_j^{'}\| \Big| \Big\langle \mathbf{v}_j^{'}, [\mathbf{U}\mathbf{x}_t]_+ \Big\rangle - \Big\langle \mathbf{v}_j', [\mathbf{U}'\mathbf{x}_t]_+ \Big\rangle \Big| = \sum_{j=1}^{h_2} \|\mathbf{w}_j'\| \Big| \Big\langle \mathbf{v}_j', [\mathbf{U}\mathbf{x}_t]_+ - [\mathbf{U}'\mathbf{x}_t]_+ \Big\rangle \Big|$$

$$\leq \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h1} |v'_{j,i}| \Big| \Big\langle \mathbf{u}_i, \mathbf{x}_t \Big\rangle - \Big\langle \mathbf{u}_i', \mathbf{x}_t \Big\rangle \Big|$$

$$\leq \sum_{j=1}^{h_2} 2\delta_j \sum_{i=1}^{h_1} |v'_{j,i}| (\alpha_i \|\mathbf{x}_t\| + \rho_{i,t}) \tag{6}$$

Thus, combining equations 1,2,3,4,5,6 we get

$$m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \frac{1}{2}\mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \sup_{\mathbf{V}'\in\Pi_V} \frac{\sqrt{2}}{\gamma} \Big[\sum_{j=1}^{h_2} 2\delta_j \sum_{i=1}^{h_1} |v_{j,i}| (\alpha_i \|\mathbf{x}_t\| + \rho_{i,t}) +$$

$$\sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} |v_{j,i} - v'_{j,i}| (\alpha_i \|\mathbf{x}_t\| + \rho_{i,t}) + \sum_{j=1}^{h_2} 2\delta_j \sum_{i=1}^{h_1} |v_{j,i}| (\alpha_i \|\mathbf{x}_t\| + \rho_{i,t}) \Big] + \tag{7}$$

$$\phi_{\mathbf{W},\mathbf{V},\mathbf{U}} + \phi_{\mathbf{W}',\mathbf{V}',\mathbf{U}'} \Big]$$

where

$$\Pi_V = \Big\{ \mathbf{V} \Big| \mathbf{V} \in \mathbb{R}^{h_2 \times h_1}, |v_{j,i} - v^0_{j,i}| \leq \beta_{j,i} \}$$

Since $\mathbf{V}$ and $\mathbf{V}'$ span the same set and using Lemma 7 from Neyshabur et al. (2018) we get:

$$m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \frac{6}{\gamma} \sum_{k=1}^{m} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \Big\langle \sigma_k, v_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k}) \Big]$$

thus proving the lemma. □

A.5 Proof of Theorem 1

Theorem 1. *Given a training set $S = \{x_i\}_{i=1}^m$ and $\gamma > 0$, Rademacher complexity of the composition of loss function l_γ over the class $\mathcal{F}_\Pi$ is bounded as follows:*

$$\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \frac{6\sqrt{2}}{\gamma\sqrt{m}} \cdot \sqrt{\sum_{j=1}^{h_2} \delta_j^2} \cdot \sqrt{\sum_{j=1}^{h_2} \sum_{i=1}^{h_1} \beta_{j,i}^2} \cdot \left(\sqrt{\sum_{i=1}^{h_1} \alpha_i^2} \frac{\|\mathbf{X}\|_F}{\sqrt{m}} + \frac{\|\mathbf{U}^0\mathbf{X}\|_F}{\sqrt{m}} \right)$$

Proof. Using Lemma 1 we have

$$m\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\Pi) \leq \mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \frac{6}{\gamma} \sum_{k=1}^{m} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \Big\langle \sigma_k, v_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k}) \Big]$$

$$\leq \frac{6}{\gamma} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \sum_{k=1}^{m} \Big\langle \sigma_k, v_{j,i} - v^0_{j,i} + v^0_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k}) \Big]$$

The previous step is an important step as it helps get rid of the terms $v^0_{j,i}$.

$$\leq \frac{6}{\gamma} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \sum_{k=1}^{m} \Big\langle \sigma_k, v_{j,i} - v^0_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k}) \Big]$$

$$+ \frac{6}{\gamma} \sum_{j=1}^{h_2} \delta_j \sum_{i=1}^{h_1} \mathbb{E}_\sigma \Big[\sup_{\mathbf{V}\in\Pi_V} \sum_{k=1}^{m} \Big\langle \sigma_k, v^0_{j,i} \Big\rangle (\alpha_i \|\mathbf{x}_k\| + \rho_{i,k}) \Big]$$

The second term in the above expression will be zero as the there is no dependence on $v_{j,i}$ except for $v_{j,i}^0$ which is a constant. Hence the expectation with respect to the Radamacher variables will be 0.

$$\leq \frac{6}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\mathbb{E}_{\sigma}\Big[\sup_{\mathbf{V}\in\Pi_V}\sum_{k=1}^{m}\Big\langle\sigma_k, v_{j,i}-v_{j,i}^0\Big\rangle(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k})\Big]$$
$$\leq \frac{6}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\mathbb{E}_{\sigma}\Big[\sup_{\mathbf{V}\in\Pi_V}\Big\langle\sum_{k=1}^{m}\sigma_k(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k}), v_{j,i}-v_{j,i}^0\Big\rangle\Big]$$
$$\leq \frac{6}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\beta_{j,i}\mathbb{E}_{\sigma}\Big[\Big|\sum_{k=1}^{m}\sigma_k(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k})\Big|\Big]$$

where the last inequality follows from the property of dual norms. Using the concavity of the square-root function we get:

$$\leq \frac{6}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\beta_{j,i}\sqrt{\mathbb{E}_{\sigma}\Big[\Big(\sum_{k=1}^{m}\sigma_k(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k})\Big)^2\Big]}$$
$$\leq \frac{6}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\beta_{j,i}\sqrt{\Big[\sum_{k=1}^{m}\Big(\alpha_i\|\mathbf{x}_k\|+\rho_{i,k}\Big)^2\Big]}$$

Using $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$

$$\leq \frac{6\sqrt{2}}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\beta_{j,i}\left(\sqrt{\sum_{k=1}^{m}(\alpha_i\|\mathbf{x}_k\|)^2}+\sqrt{\sum_{k=1}^{m}(\rho_{i,k})^2}\right)$$
$$\leq \frac{6\sqrt{2}}{\gamma}\sum_{j=1}^{h_2}\delta_j\sum_{i=1}^{h_1}\beta_{j,i}\Big(\alpha_i\|\mathbf{X}\|_F+\|\mathbf{u}_i^0\mathbf{X}\|\Big)$$

Using Cauchy-Schwarz, two times we get:

$$m\mathcal{R}_S(\ell_\gamma\circ\mathcal{F}_\Pi) \leq \frac{6\sqrt{2}}{\gamma}\cdot \underbrace{\sqrt{\sum_{j=1}^{h_2}\delta_j^2}}_{\text{upper bound on}\|\mathbf{W}\|_F} \cdot \underbrace{\sqrt{\sum_{j=1}^{h_2}\sum_{i=1}^{h_1}\beta_{j,i}^2}}_{\text{upper bound on}\|\mathbf{V}-\mathbf{V}^0\|_F} \cdot \underbrace{\left(\sqrt{\sum_{i=1}^{h_1}\alpha_i^2}\|\mathbf{X}\|_F+\|\mathbf{U}^0\mathbf{X}\|_F\right)}_{\text{upper bound on}\|\mathbf{U}-\mathbf{U}^0\|_F} \tag{8}$$

Thus concluding the proof. This upper bound decreases with increasing the number of hidden units in the neural network and hence is consistent with our observations. □