Nonvacuous Generalization Bounds For Deep Networks With Improved Size Dependence

Anonymous authors
Paper under double-blind review

Abstract

Despite being massively overparameterized, deep neural networks exhibit a remarkable ability to generalize well to unseen data. Existing generalization bounds fail to explain this phenomenon, often becoming vacuous due to their strong dependence on network depth and width. To address this, we introduce novel nonvacuous generalization bounds for deep networks, offering tighter estimates of their Rademacher complexity by introducing a new analysis of covering number, which exhibits much milder depth dependence. Our bounds grow at a much slower rate of $O(\sqrt{Dpr})$, with network depth D, width p, and weights of rank r, compared to previous works that scale at a rate $O(\sqrt{D^3pr})$. Moreover, under certain plausible assumptions on the norms of network weights, we establish bounds that grow at a sub-logarithmic rate of $O(\sqrt{log\ D})$ with depth. This novel bound is much tighter and represents a substantial improvement over prior bounds that scale at a polynomial rate with depth. We provide rigorous empirical validation, demonstrating that our bounds offer consistently tighter estimates compared to the state-of-the-art results. Thus, our bounds offer improved insight into the excellent generalization capabilities of deep overparameterized networks.

1 Introduction

Deep learning models have achieved remarkable success in many domains including computer vision, Natural Language Processing, and so on. Yet, a fundamental theoretical puzzle remains: Why do over-parameterized deep networks, with many more parameters than training data, still generalize well to unseen data (Zhang et al., 2021)? Providing a theoretical justification for this generalization puzzle has remained elusive.

Generalization bounds are a powerful tool for assessing the generalization performance of deep neural networks. Traditional learning theory offers bounds, such as those based on the VC dimension (Vapnik & Chervonenkis, 2015) that are derived from the principle of uniform convergence (Nagarajan & Kolter, 2021). However, these bounds typically grow in proportion to the number of parameters in the network. For massive neural networks, these bounds become *vacuous* and cannot truly capture the generalization gap. Thus, these bounds are of little use in explaining the good generalization behavior of very large networks because of their explicit dependence on the size of these networks.

To address these limitations, researchers have explored alternative measures, such as the *norm* of the network weights (Neyshabur et al., 2015) and the classifier *margin* (Bartlett et al., 2017), to provide tighter bounds that are nonvacuous, more conservative, and correlate better with the generalization gap compared to those based on traditional methods.

Although bounds based on these measures have demonstrated a significant improvement, a new challenge has emerged: even these tighter bounds scale poorly with increasing depth and width of the network, whereas empirical observation suggests that the generalization gap of a network is almost independent of its width and grows slowly with its depth. For instance, Neyshabur et al. (2015) defines bounds, based on weight norms, that grow exponentially with the depth D of the network (at a rate $O(2^D)$) and are thus vacuous, despite being tighter than traditional bounds based on the VC dimension. Similarly, Bartlett et al. (2017) and Neyshabur et al. (2018) use the concept of covering numbers to suggest bounds that grow at a fast

polynomial rate $O(\sqrt{D^3pr})$ with the network dimensions, where D denotes depth, p the maximum number of hidden units in any layer, and r denotes the rank of the weight matrices, even in special cases of bounded norms in each layer. However, these bounds still grow at a faster rate with the depth of the network, which might lead to loose bounds.

To address their shortcomings, we propose a novel generalization bound that exhibits a significantly milder dependence on the network dimensions. Our bounds grow only at a rate $O(\sqrt{Dpr})$, and in the special case of a bounded norm of the weights of hidden units in each layer, the dependence becomes even more conservative, growing only sub-logarithmically with depth at a rate $O(\sqrt{log\ D})$. Thus, our bounds are much tighter, nonvacuous, growing at a slower pace with increasing network dimensions, and thus provide a better theoretical explanation for the good generalization performance of over-parameterized networks.

To provide a detailed derivation of the bounds and show that it captures the generalization gap well, we have organized the paper as follows. In Section 2, we present related works that attempted to provide nonvacuous generalization bounds for deep networks to provide context for our work. In Section 3, we provide a rigorous derivation of our generalization bounds, starting with the general condition and then proving it for the special norm-constrained case. We then compare our bounds with related bounds on the generalization gap of deep networks. In Section 4, we experimentally evaluate our bounds and show that they are tighter than the related bounds suggested previously. In Section 5, we provide a detailed discussion of the empirical comparison of the different bounds evaluated on networks with different depths. Finally, we conclude in Section 6 while suggesting interesting directions for future research in Section 7.

In general, our work makes the following contributions:

- We propose a novel generalization bound that is nonvacuous and has a better dependence on the network dimensions (depth and width) than previously suggested bounds. This marks a step forward in obtaining tighter bounds that scale well with the network size.
- Under certain realistic and practical assumptions on the norm of the network weights, we provide bounds that scale only sub-logarithmically with a rate $O(\sqrt{log\ D})$ with depth D, which is slower than the prior works that scale at a rate of at least $O(\sqrt{D})$. Crucially, our assumptions are more broadly applicable and more plausible than the restrictive ones made in works such as (Golowich et al., 2018).
- Through extensive experiments, we provide empirical evidence demonstrating the tightness of our bound and its superior dependence on depth. By demonstrating the significantly lower values yielded by our bound over networks of increasing depth compared to competitive baselines, we confirm that our nonvacuous bounds are much tighter and scale better with depth, providing a superior alternative to the prior bounds.

1.1 Preliminaries

This section defines key concepts related to DNNs, loss functions, and generalization bounds. The key terms defined here will be essential for the subsequent analysis.

Given an input x, we denote the output of a D layer neural network as $f(x) = f_{w_d}(f_{w_{d-1}}(..f_{w_1}(x)))$, where $w_d, w_{d-1}, .., w_1$ denote the weights of this network in layers $d, d-1, ..., 1$. The output of any intermediate layer d is given as $f_{w_d}(y_{d-1}) = \sigma(W_d y_{d-1})$, where $\sigma(.)$ is the activation function which, in our case, is the ReLU activation (Agarap, 2019) and y_d denotes the output of the d_{th} layer.

We denote loss functions as $g(f(x), y)$ where the first argument specifies the network output corresponding to the input x, and the second specifies the target y. We then denote a class of loss functions defined over the function class $\mathcal{F}_D$ representing D-layer neural networks as

$$\mathcal{G}(\mathcal{F}_\mathcal{D}) = \{g(f(w, x_i), y_i) : w \in \mathcal{W}, (x_i, y_i) \in (\mathcal{X}, \mathcal{Y})\} \tag{1}$$

where $\mathcal{W}$ specifies the set representing all plausible values of the network weights and $\mathcal{X}, \mathcal{Y}$ denote the sets of possible inputs and their corresponding observations, respectively. Given a set of m inputs

$(x_1, y_1), (x_2, y_2), ..(x_m, y_m)$, the generalization error of a classifier f is the difference between its true (or expected) risk and its empirical risk on this subset of size m and can be written as:

$$\mathbb{P}[arg\ \underset{i}{max}(f(x)_i) \neq y] - \frac{1}{m}\sum_{i=1} g(f(x_i), y_i) \tag{2}$$

Let us take the task of classification as an example to elucidate this concept further. Specifically, we consider the multiclass classification with c classes. Consider a particular bounded form of the loss function $g(.)$, i.e., the $margin - loss$ or $ramp - risk$. For a given input x with the correct class $y \in \{1, 2, .., c\}$, let us define $m_W^{x,y} = f_y(x) - \underset{i \neq y}{max} f_i(x)$. We can then define the class of *margin loss* functions with a margin γ as:

$$\mathcal{G}_\gamma(\mathcal{F}_D) = \{g_\gamma(f(w, x), y) : \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall f \in \mathcal{F}_D, \forall w \in \mathcal{W}\} \tag{3}$$

where

$$g_\gamma(f(x), y) = \begin{cases} 0, & m_W^{x,y} > \gamma \\ 1 - \frac{m_W^{x,y}}{\gamma}, & 0 \leq m_W^{x,y} \leq \gamma \\ 1, & m_W^{x,y} < 0 \end{cases} \tag{4}$$

Thus, the margin-loss being 1 indicates an incorrect prediction, while it is 0 when the model output for the correct class is greater than that of all other classes by the margin γ. Otherwise, if the output corresponding to the correct class is greater than that of the maximum of other classes by a value $\leq \gamma$, this loss goes down from 1 towards 0 as this difference reaches γ.

After defining these crucial concepts, we now define some important terms for understanding generalization bounds. The empirical Rademacher complexity (Bartlett & Mendelson, 2002) $\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_D))$ for a sample set of size n over a function class $\mathcal{F}_D$ is described as

$$\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) = \underset{\epsilon_i \in [-1,+1]^n}{\mathbb{E}}[\ \underset{f \in \mathcal{F}_D}{sup}\ |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(f(x_i), y_i)|] \tag{5}$$

where ϵ is a vector that contains -1 or +1 with equal probability. This quantity will play a crucial role in the derivation of our generalization bounds.

Bartlett et al. (2017) connects the generalization gap to the Radamechar complexity as follows: Given a function class representing D-layer neural networks $\mathcal{F}_D$, a sample set S of size n, margin loss function g_γ, and a real number $\delta \in [0, 1]$, we have, with probability $> 1 - \delta$, that the generalization error is bounded with respect to the Rademacher complexity via the following inequality:

$$\mathbb{P}[\underset{i}{argmax}(f(x)_i) \neq y] \leq \frac{1}{n}\sum_{i=1}^{n} g_\gamma(f(x_i), y_i) + 2\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) + 3\sqrt{ln\frac{2/\delta}{n}} \tag{6}$$

1.2 Notations

Now we state some important notations that would be regularly used in the mathematical expressions below. We will usually use D, p, and r to denote the depth, maximum width, and the rank of the weight matrices of a neural network. Given a matrix W_d, which represents the weights of the d_{th} layer of a neural network, we denote its spectral norm as $B_{d,2}$ or $\|W_d\|_2$, the Frobenius norm as $B_{d,F}$ or $\|W_d\|_F$, and $B_{d,2\to1}$ or $\|W_d\|_{2,1}$ denotes the 2,1-norm of the d_{th} layer, which is described mathematically as $\|W\|_{2,1} = \sum_{i=0}^{n}\sqrt{\sum_{j=1}^{m} W_{ij}^2}$. This is just the sum of the l_2 norm of each row of the matrix W_d. For a given neural network, J_{W_d} will denote the Jacobian of the d_{th} layer, $B_{1:d}^{jac}$ denotes the Jacobian norm of the first d layers with respect to the input and $B_{\setminus d}^{jac}$ denotes the Jacobian of the full network excluding the d_{th} layer. Furthermore, γ denotes the classifier margin. R will denote the L_2 norm of the input matrices $\|X\|_2$, i.e., $R = \|X\|_2 = \sum_{i=1}^{m}\sqrt{\|x_i\|_2^2}$. $\rho_1, \rho_2, .., \rho_D$ denotes the Lipschitz constant of the respective layers in a D-layer neural network

and $\mathcal{N}(\mathcal{F}, \epsilon, \|.\|_p)$ denotes the covering number of the function class $\mathcal{F}$ learnable by the neural network. This covering number determines the number of balls of size ϵ required to cover the whole class of neural network functions represented by $\mathcal{F}$, with the distance measured in the $\|.\|_p$ norm.

2 Related Works

2.1 Nonvacuous generalization bounds for deep networks

Traditional generalization bounds for neural networks, such as those based on the VC dimension (Vapnik & Chervonenkis, 2015) scale linearly with the number of network parameters. Although theoretically grounded in statistical learning theory, these bounds are vacuous for overparameterized neural networks, thus providing little insight into the generalization performance of deep networks. A recent line of work has attempted to address this limitation by proposing nonvacuous generalization bounds that are some functions of the *norm* of the weights (Bartlett et al., 2017; Neyshabur et al., 2018; 2015) and the classifier *margin*. Unlike traditional bounds, these methods offer more favorable scaling and better estimation of the generalization gap.

Early works in this direction (Neyshabur et al., 2015) proposed bounds that are explicitly based on the $l_{p,q}$ norm of the network weights. In this approach, the l_p norm of the weights incoming to each hidden unit is first computed, which is then followed by a l_q norm of the resulting vector. While this bound is a notable improvement over those based on the VC dimension, it still suffers from a critical limitation. This bound scales exponentially with the depth of the network (scaling at a rate $O(\frac{2^D \prod_d B_{d,F}}{\gamma\sqrt{m}})$, where D is the depth), which does not align well with empirical observations according to which the generalization gap grows slowly and often improves with depth (Nakkiran et al., 2020).

One of the seminal works that addressed the fast scaling of the previous bounds with depth is the one proposed by Bartlett et al. (2017). This bound leverages the concept of covering number for neural networks to derive nonvacuous bounds that are proportional to the $l_{2,3}$ norm of the ratio of the Frobenius norm to the spectral norm of the weights across the layers to provide tighter generalization estimates.

$$\frac{\prod_{d=1}^{D} B_{d,2} log\ p}{\gamma\sqrt{m}} \sqrt{\Big(\sum_{d=1}^{D} \frac{B_{d,2\to1}^{2/3}}{B_d^{2/3}}\Big)^3} \tag{7}$$

However, their bound grows at a rate of at least $O(\frac{\sqrt{D^3pr}}{\sqrt{m}})$, which is proportional to the square root of the cube of the network depth. Although their bounds showed an improved growth rate with network dimensions compared to (Neyshabur et al., 2015), it still did not result in bounds with a tighter dependence on depth.

Neyshabur et al. (2018) later proposed a tighter bound that is based on PAC-Bayesian principles (McAllester, 2003) and is proportional to the sum of the ratio of the squares of the Frobenius norm to the spectral norm across all layers multiplied by D^2p.

$$\frac{\prod_{d=1}^{D} B_{d,2}\ log(Dp)}{\gamma\sqrt{m}} \sqrt{D^2p\left(\sum_{d=1}^{D} \frac{B_{d,F}^2}{B_{d,2}^2}\right)} \tag{8}$$

Despite being an improvement over the previous bound, this bound still grows at the rate $\frac{\sqrt{D^3pr}}{\sqrt{m}}$ due to the extra D^2p term, and thus scales at the same rate as the previous bound, since the sum of the ratio of the Frobenius norm to the spectral norm scales as $O(D)$. Due to the fast growth rate of these bounds with network depth, which contradicts empirical findings, these bounds cannot provide an accurate estimate of the generalization gap, especially for very deep networks.

The fast growth rates of the previous bounds present a significant challenge in providing tight and nonvacuous estimates of the generalization gap of the network. This can be a significant hurdle in providing an estimate that is closer to the actual generalization gap of very deep networks, as these bounds grow very fast with network depths, while empirical observation suggests that the generalization error grows more

slowly with depth (slower than a polynomial rate), and even adding depth under certain conditions improves generalization (Nakkiran et al., 2020; Advani & Saxe, 2017).

To address the limitations of previous bounds and provide better depth scaling, Li et al. (2019) provided a tighter bound that scales as $O(\frac{\sqrt{Dpr*log\ C_{net}}}{\gamma\sqrt{m}})$, where C_{net} is a term that is proportional to the square root depth of the network and its Jacobian. Thus, their bound scale as $O(\sqrt{Dpr*log(c\sqrt{D})})$, where c is a constant, which is tighter compared to the previous bounds. A more recent attempt at obtaining size-independent bounds was made in (Golowich et al., 2018), where the authors propose a bound which is proportional to the logarithm of the ratio of the product of the Frobenius norm of the layer weights and its lower bound $\Gamma \leq \prod_{d=1}^{D} B_{d,F}$:

$$\prod_{d=1}^{D} \frac{B_{d,F}}{\gamma}\ min\left\{\frac{1}{\sqrt[4]{m}}\sqrt{log\ \frac{\prod_{d=1}^{D} B_{d,F}}{\Gamma}}, \sqrt{\frac{D}{m}}\right\} \tag{9}$$

However, their bounds only hold in a very stringent condition where the Frobenius norms of the weights in each layer d are upper-bounded by some constant $B_{d,F}$, while their product is lower-bounded by Γ. Moreover, they used the product of the Frobenius norm of the layers, instead of the tighter spectral norm used by prior bounds, which can lead to terms that scale exponentially with depth ($O(r^D)$) in the worst case, resulting in loose bounds. Lastly, their bound is independent only in the case where the first term in the min operation ($\sqrt{\frac{1}{\sqrt[4]{m}}log\ \frac{\prod_{d=1}^{D} B_{d,F}}{\Gamma}}$) is sufficiently smaller than the second term $\sqrt{D/m}$, which is more rarely satisfied due to an almost exponential growth of $B_{d,F}$ with depth.

Compared to previous works, we propose bounds that exhibit a significantly better dependence on network dimensions (both depth and width). Our bounds scale as $O(\frac{\sqrt{Dpr}}{\sqrt{m}})$, and thus have a better depth scaling than all the previous bounds, including (Li et al., 2019) which scaled as $O(\sqrt{\frac{Dpr*log(D)}{m}})$.

Moreover, under certain assumptions on the norm of the weights of the network layers, our bounds achieve a remarkable sub-logarithmic scaling with depth at a rate $\left(\frac{\sqrt{log(D)}}{\sqrt{m}}\right)$. This provides a significantly tighter and realistic estimate of the generalization gap of deeper neural networks with many more layers, addressing a major limitation of previous work and aligning well with the empirical observation.

3 An improved generalization bound with better size dependence

This section presents our novel generalization bounds and provides rigorous mathematical derivations for them. We begin by establishing our bounds in the general case without any restrictions on the network norm. We then demonstrate that under certain practical constraints on the norm of the network's weights, a tighter generalization bound can be obtained, which scales only sub-logarithmically with the depth D of the network and has a tighter dependence on both depth and width.

3.1 The bound in the general case

To derive this bound, we first propose the following upper bound on the covering number of the function class of a deep neural network with D layers, which will serve as a key component in our analysis.

Lemma 1 *Let $\mathcal{F}_D$ denote the function class represented by a deep neural network with D layers, whose weights W_i in each layer $i \in \{1, 2, .., D\}$ satisfy $\|W_i - W_{i_0}\|_{2,1} \leq b_i$, where W_{i_0} denotes the weights at initialization. Furthermore, if the spectral norm of the weights of i_{th} layer satisfy $\|W_i\|_2 \leq s_i$ for all $i \in \{1, 2, .., D\}$, then the log of the covering number of the function class $\mathcal{F}_D$ under the L_2-norm is bounded by:*

$$ln\ \mathcal{N}(\mathcal{F}_D, \epsilon_i, \|.\|_2) \leq \Big(\prod_{i=1}^{D} \rho_i s_i\Big)\Big(\sum_{i=1}^{D}\Big(\frac{b_i}{s_i}\Big)^2\Big)^{1/2} = \mathcal{R}_D \tag{10}$$

where $\mathcal{R}_D$ is defined as the spectral complexity of the network.

This covering number bound serves as an essential component for bounding the spectral complexity of the network, denoted by $\mathcal{R}_D$, which will be useful in deriving our generalization bounds. Here, we deviate from previous work (Bartlett et al., 2017) by replacing the arbitrary reference matrices M_i with matrices representing the weights at initialization W_{i_0}, to obtain tighter bounds. This choice was informed by the empirical observation that the weights of the neural network tend to remain closer to their initialization throughout the training phase, and also theoretically explained in (Nagarajan & Kolter, 2019). Based on this, we propose our novel margin-based generalization bound in the theorem below.

Theorem 2 *Let us assume a D-layer neural network with weights in each layer given by $W_1, W_2, ...W_i, .., W_D$ and these weights during initialization are given as $W_{1_0}, W_{2_0}, W_{3_0}, .., W_{i_0}, ..W_{D_0}$. If $(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_m, y_m)$ are drawn iid from a distribution, then with probability at least $1 - \delta$, the true error rate is bounded by (Bartlett et al., 2017):*

$$\mathbb{P}[arg\ \underset{i}{max}(f(x)_i) \neq y_i] \leq \hat{\mathcal{R}}(f) + O\Big(\frac{\|X\|_2 \mathcal{R}_D}{\gamma m} ln(K) + \sqrt{ln\frac{1/\delta}{n}}\Big) \tag{11}$$

where γ denotes the margin of the network, $\mathcal{R}_D$ the spectral complexity (Bartlett et al., 2017), $\hat{\mathcal{R}}(f) = m^{-1}\sum_{i=1}^{m} \mathbb{I}[f(x_i)_{y_i} \leq \underset{j \neq y_i}{max}\ f(x_i)_j + \gamma]$ is the empirical risk, K= $max(d_0, d_1, d_2, ..., d_D)$ where d_i denotes the dimension of the i_{th} layer output, and $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

Substituting the value of the spectral complexity ($\mathcal{R}_D$) derived in Lemma 1 and replacing $b_i = \|W_i - W_{i_0}\|_{2,1}$ and $s_i = \|W_i\|_2$, this simplifies to our refined bound:

$$\mathbb{P}[arg\ \underset{i}{max} f(x)_i \neq y] \leq \hat{\mathcal{R}}(f) + O\left(\frac{\|X\|_2 \prod_{i=1}^{D} \rho_i \|W_i\|_2 ln(K^2)}{\gamma m} \sqrt{\sum_{i=1}^{D} \frac{\|W_i - W_{i_0}\|_{2,1}^2}{\|W_i\|_2^2}} + \sqrt{ln\frac{1/\delta}{m}}\right)$$

The full derivation of this theorem is described in the Appendix. The above theorem states our generalization bound for the general condition. Note that our bound on the covering number is of the order $O(\sqrt{Dpr})$, which is much tighter than the bound proposed in (Bartlett et al., 2017) which is based on the following upper bound on the covering number of the network:

$$\mathfrak{R}_n(\mathcal{F}_D) \leq \sqrt{ln\ \mathcal{N}(\mathcal{F}, \epsilon_i, \|.\|_\infty)} < \prod_{i=1}^{D} \rho_i \|W_i\|_2 \sqrt{\Big(\sum_{i=1}^{D} \Big(\frac{\|W_i\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\Big)\Big)^3} \tag{12}$$

that grows at a rate $O(\sqrt{D^3pr})$. Moreover, incorporating the distance of the weights from initialization in our measures leads to tighter bounds that align well with the empirical observation suggesting that the weights stay consistently closer to their initialization throughout training (Nagarajan & Kolter, 2019).

3.2 The bound in the norm-constrained case

Now, we derive our bound for the special case where we assume that the norms of the weights of each hidden unit is bounded by a constant $C = 1$ (i.e., $\|W_{ik}\|_2 < 1$ for each layer i and hidden unit k). Under this assumption, we will use Jensen's inequality to obtain a tighter bound, which only scales sub-logarithmically with depth D.

To derive this bound, we leverage the following well-known result from statistical learning theory that bounds the Rademacher complexity (Bartlett & Mendelson, 2003) $\mathfrak{R}_n(\mathcal{F})$ of a function class $\mathcal{F}$ using the covering number:

$$\mathfrak{R}_n(\mathcal{F}) = \underset{\beta > 0}{inf}\left(\beta + \int_{\beta}^{B} \sqrt{log\ \mathcal{N}(\mathcal{F}, \epsilon, L_2(\mathcal{F}))}\right) \tag{13}$$

where $\mathcal{N}(\mathcal{F}, \epsilon, L_2(\mathcal{F}))$ denotes the ϵ-covering of the neural network function class $\mathcal{F}$ with respect to the metric $L_2(\mathcal{F})$ in the function class $\mathcal{F}$. Now, based on the assumptions on the network's weights, we derive an upper bound on the Rademacher complexity in the following theorem.

Theorem 3 *Let us assume a D-layer neural network, with at most p hidden units in any layer. We assume that the weights in each layer have a maximum rank of r and are denoted by $W_1, W_2, ...W_i, .., W_D$, and the weights during initialization are given as $W_{1_0}, W_{2_0}, W_{3_0}, .., W_{i_0}, .., W_{D_0}$. Further, let us assume that the norm of each hidden unit is upper bounded by a constant $C(= 1)$ (i.e., $\|W_{ij}\|_2 \leq 1$).*

Let $S = (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_m, y_m)$ be a set of m training examples drawn iid from a distribution. Let $\mathcal{F}_D$ denote the function class of all such neural networks. Then, with probability at least $1 - \delta$ over a random draw of the set S, the Rademacher complexity of the margin loss function class $\mathcal{G}_\gamma$ over the network function class $\mathcal{F}_D$ is bounded as

$$\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) = B^{jac}_{1:D} R\left(\sqrt{\frac{2\sqrt{2}p^{3/2}r}{\gamma B^{jac}_{1:D}}} + \frac{1}{\gamma\sqrt{m}}\sqrt{log\Big(\sum_{i=1}^{D}\frac{\|W_i - W_{i_0}\|_{2,1}}{\|W_i\|_2\sqrt{2p}}\Big)}\right) \tag{14}$$

The detailed proof of this theorem is relegated to the Appendix.

Now, utilizing the fact that the spectral norm is bounded by the Frobenius norm (i.e., $\|W_i\|_2 \leq \|W_i\|_F$) and using the assumption that the norm of each hidden unit weight is constrained by $C = 1$, we can show that for the i_{th} layer:

$$\|W_i\|_2 \leq \|W_i\|_F = \sqrt{\sum_{j=1}^{p}\sum_{k}(W_{ijk})^2} = \sqrt{\sum_{j=1}^{p}\|W_{ij}\|_2^2} \overset{(i)}{=\leq} \sqrt{\sum_{i=1}^{p} 1} \leq \sqrt{p} \approx O(\sqrt{p})$$

where inequality (i) follows from the assumption that $\|W_{ij}\|_2 < 1$.

Moreover, the numerator in the sum in Equation (14) scales as

$$\|W_i - W_{i_0}\|_{2,1} = \sum_{j=1}^{p}\|W_{ij} - W_{ij_0}\|_2 \leq \sum_{j=1}^{p}(\|W_{ij}\|_2 + \|W_{ij_0}\|_2) \leq \sum_{i=1}^{p} 2 \leq 2p$$

Since $\|W - W_{i_0}\|_{2,1}$ and $\|W_i\|_2$ scale as $O(2p)$ and $O(\sqrt{p})$, individual terms within the summation in the log in Theorem 3 scale as $O(2p/\sqrt{p} \times \sqrt{2p}) \approx O(1)$, hence the value of the summation inside the log in (14) can grow at a maximum rate of $O(D)$ and the Rademachar complexity term derived in the theorem scales at a rate $O(\sqrt{\sqrt{p^{3/2}}r/B^{jac}_{1:D}} + \sqrt{log\ D})$. Thus, our bound scales only sub-logarithmically with respect to the depth D.

Furthermore, since the spectral norm of each layer with p units grows as $\|W_d\|_2 \leq O(\sqrt{p})$, the Jacobian of the network can be shown to grow as $B^{jac}_{1:D} = \prod_d B_{d,2} = \prod_d \|W_d\|_2 \leq O((\sqrt{p})^D)$. Thus, the overall generalization bound should grow at a rate $O(\sqrt{p^{3/2}r/p^D} + \sqrt{log\ D}) \approx O(\sqrt{log\ D} + \sqrt{r/p^{(D-1.5)}})$ with respect to the network dimensions. Hence, on the basis of this observation, our proposed bound should grow only sub-logarithmically with depth and should even decrease with increasing width of the network for constant depth. Now, substituting (14) into the following generalization bound (Bartlett et al., 2017) based on the Rademacher complexity

$$\mathbb{P}[arg\ \underset{i}{max}(f(x)_i) \neq y_i)] \leq \hat{\mathcal{R}}(f) + 2\mathfrak{R}_n(\mathcal{F}) + 3\sqrt{ln\frac{1/\delta}{n}} \tag{15}$$

we finally get our refined generalization bound.

Theorem 4 *Let's assume a D layer neural network, with a maximum of p hidden units in any layer, weights in each layer given by $W_1, W_2, ...W_i, .., W_D$ and having a maximum rank r, and the corresponding weights*

Table 1: Comparison of existing norm-based generalization bounds with our results for neural networks of depth D, p hidden units in each layer, and weight matrices of rank r. $B_{d,F}, B_{d,2}$, and $B_{d,2\to1}$ denote the $\|W_d\|_F, \|W_d\|_2, \|W_d - W_{d_0}\|_{2,1}$ norms and $\Gamma \leq \prod_{d=1}^{D} \|W_d\|_2$ denotes the lower bound on product of Frobenius norm (Golowich et al., 2018). We also derive each bound under the constrain $\|W_d\|_2 = 1$ for all the layers $d = 1, 2, .., D$ which makes the common factor $\prod_{d=1}^{D} B_{d,2}$=1.

capacity bound	Original results	$\|W_d\|_2 = 1$
(Neyshabur et al., 2015)	$\mathcal{O}\Big(\frac{2^D \prod_{i=1}^{D} B_{d,F}}{\gamma\sqrt{m}}\Big)$	$\mathcal{O}\Big(\frac{2^D r^{D/2}}{\gamma\sqrt{m}}\Big)$
(Bartlett et al., 2017)	$\mathcal{O}\Big(\frac{\prod_{d=1}^{D} B_{d,2} R\ log\ p}{\gamma\sqrt{m}}\Big(\sum_{l=1}^{D} \frac{B_{d,2\to1}^{2/3}}{B_d^{2/3}}\Big)^{3/2}\Big)$	$\widehat{\mathcal{O}}(\frac{\sqrt{D^3pr}}{\gamma\sqrt{m}})$
(Neyshabur et al., 2018)	$\mathcal{O}\Big(\frac{\prod_{d=1}^{D} B_{d,2} R\ log(Dp)}{\gamma\sqrt{m}}\sqrt{D^2p\sum_{d=1}^{D}\frac{B_{d,F}^2}{B_{d,2}^2}}\Big)$	$\widehat{\mathcal{O}}(\sqrt{\frac{D^3pr}{\gamma\sqrt{m}}})$
(Golowich et al., 2018)	$\mathcal{O}\Big(\prod_{d=1}^{D}\frac{B_{d,F}}{\gamma}\ min\Big\{\frac{\sqrt{log\frac{\prod_{d=1}^{D} B_{d,F}}{\Gamma}}}{m^{1/4}}, \sqrt{\frac{D}{m}}\Big\}\Big)$	$\mathcal{O}\Big(\frac{\sqrt{r^D D}}{\gamma\sqrt[4]{m}}\Big)$
(Li et al., 2019)	$\mathcal{O}\Big(\frac{B_{1:d}^{jac}\sqrt{Dpr}\ log\Big(\frac{B_{\setminus d,2}^{jac}\sqrt{Dm/r}\ max_x B_{d,2}}{sup\ g(f(\mathcal{W}_D,x))}\Big)}{\gamma\sqrt{m}}\Big)$	$\mathcal{O}\Big(\frac{\sqrt{Dpr*log(D)}}{\gamma\sqrt{m}}\Big)$
Our bound	$\mathcal{O}\Big(\frac{\prod_{d=1}^{D} B_{d,2} R\ log(p)}{\gamma\sqrt{m}}\sqrt{\sum_{d=1}^{D}\frac{B_{d,2\to1}^2}{B_{d,2}^2}}\Big)$	$\widehat{\mathcal{O}}(\frac{\sqrt{Dpr}}{\gamma\sqrt{m}}$
Our bound (norm-constrained)	$\mathcal{O}\Big(\frac{\prod_{d=1}^{D} B_{d,2} R}{\gamma\sqrt{m}}\sqrt{log\sum_{d=1}^{D}\frac{B_{d,2\to1}}{B_{d,2}\sqrt{p}}}\Big)$	$\widehat{\mathcal{O}}(\frac{\sqrt{log(D)}}{\gamma\sqrt{m}} + \sqrt{\frac{r}{\gamma\ p^{D-1.5}\sqrt{m}}})$

during initialization are given as $W_{1_0}, W_{2_0}, W_{3_0}, .., W_{i_0}, .., W_{D_0}$. If $(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)$ is a sample S of size n drawn iid from a distribution and assuming that the norm of each hidden unit's weight is bounded above by 1, then with probability at least $1 - \delta$, over random draw of S, the generalization error is bounded as

$$\mathbb{P}[\underset{i}{argmax}(f(x)_i) \neq y] \leq \hat{\mathcal{R}}(f) + O\Big(\frac{\|X\|_2 B_{1:d}^{jac}}{\gamma n}\sqrt{log\ \sum_{i=1}^{D}\frac{\|W_i - W_{i_0}\|_{2,1}}{\|W_i\|_2\sqrt{2p}}} + \frac{Cp^{3/4}r^{1/2}}{\sqrt{B_{1:D}^{jac}}} + \sqrt{ln\frac{2/\delta}{n}}\Big) \quad (16)$$

where γ denotes the classifier margin, $\hat{R}(f) = \sum_{i=1}^{n}\frac{(argmax_j f(x_i)_j \neq y_i)}{n}$ is the empirical risk and $\|X\|_2 = \sqrt{\sum\|x_i\|_2}$.

Note that our bound is tighter than Bartlett et al. (2017) as well as the more recent bound Li et al. (2019), and scales only sub-logarithmically with depth. We now provide an empirical comparison of the growth rate and the tightness of our bound compared to the other related bounds.

4 Experimental Evaluation

We evaluated and compared our generalization bound against the ones proposed in Bartlett et al. (2017); Neyshabur et al. (2018); Li et al. (2019)and Golowich et al. (2018) on a VGG-16 network and its deeper variants (VGG-19, VGG-21, and VGG-26) on the CIFAR-10 dataset (Krizhevsky, 2009). VGG-21 and VGG-26 models are extensions of VGG-19, where VGG-21 had one more convolutional layer in the last two blocks, each consisting of 512 filter layers. Similarly, VGG-26 was an extension of VGG-19 and had 3 more layers in the last two convolutional blocks with 512 filter layers in each block, and 1 more convolutional layer in the third block consisting of 256 filter layers.

In evaluating these bounds, we assumed that the common factor $\frac{\|X\|_2}{\gamma\sqrt{m}}$ was the same and equal to 1 for a specific network across all bounds. We evaluated the following bounds after training the networks on the CIFAR-10 dataset (Krizhevsky, 2009) to assess their tightness and overall performance:

- $\prod_{i=1}^{D} \|W_i\|_2 \sqrt{\sum_{i=1}^{D} \frac{\|W_i\|_{2,1}^2}{\|W_i\|_2^2}}$, Bound 1a (ours, General)
- $\prod_{i=1}^{D} \|W_i\|_2 \sqrt{\sum_{i=1}^{D} \frac{\|W_i - W_{i_0}\|_{2,1}^2}{\|W_i\|_2^2}}$, Bound 1b (ours, Initialization based)
- $\prod_{i=1}^{D} \|W_i\|_2 \sqrt{log\Big(\sum_{i=1}^{D} \frac{\|W_i - W_{i_0}\|_{2,1}}{\|W_i\|_2 \sqrt{p}}\Big)}$, Bound 1c (ours, Norm-constrained)
- $\prod_{i=1}^{D} \|W_i\|_2 * \sqrt{k \sum_{i=1}^{D} n_i}$, Bound 2 (Tighter bound proposed in (Li et al., 2019))
- $\prod_{i=1}^{D} \|W_i\|_2 \sqrt{\Big(\sum_{i=1}^{D} \frac{\|W_i\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\Big)^3}$, Bound 3 (spectrally normalized margin bound of (Bartlett et al., 2017))
- $\prod_{i=1}^{D} \|W_i\|_2 \sqrt{D^2 p\Big(\sum_{i=1}^{D} \frac{\|W_i\|_F^2}{\|W_i\|_2^2}\Big)}$, Bound 4 (PAC-Bayesian bound proposed in (Neyshabur et al., 2018))
- $\prod_{i=1}^{D} \|W_i\|_F * min\{log(\frac{\prod_{i=1}^{D} \|W_i\|_F}{\Gamma}), \sqrt{D}\} \approx \prod_{i=1}^{D} \|W_i\|_F \sqrt{D}$, Bound 5 (size-independent bound of (Golowich et al., 2018))

The values of these bounds for the different variants of the VGG networks are documented in Table 2.

We initialized the weights of the networks using Glorot-uniform initialization (Glorot & Bengio, 2010). During training, we constrained the filters to be of unit norm in all the models, following Li et al. (2019) and to satisfy the constraint in Theorem 4. We trained the network for 60 epochs with the SGD optimizer with 0.9 momentum and an initial learning rate of 0.01, which was decayed by half after the first 20 epochs, followed by the same decay of 0.5 every 10 epochs thereafter.

After an evaluation of the bounds for networks of different depths, we found that our bound in the general case (Bound 1a) was of order 10^5 for the VGG-16 network, whereas bounds 3 and 4 were of order 10^8 and 10^7, respectively, while Bound 5 (Golowich et al., 2018) was of order 10^{17} which was extremely loose due to its dependence on product of Frobenius norm across layers instead of the spectral norm. Our bounds were tighter than even the recently proposed bound by Li et al. (2019), which was of the order of 10^6 for the same network, despite its more favorable size dependence. Furthermore, our bound based on the distance from initialization (Bound 1b) of the weights was even tighter than our general Bound 1a on the same network. This demonstrates that the distance of weights from initialization can provide a tighter characterization of the generalization gap compared to regular norms. In addition, our sub-logarithmic bound (Bound 1c) was even tighter and was of the order 10^4.

Similarly, on a VGG-21 network, our general bound (Bound 1a) was of order 10^7, while Bounds 3 and 4 were of order 10^{10} and 10^9, and Bound 5 was of order 10^{24}. Moreover, on the same network, our Bound 1b was again almost half of that of Bound 1a, and our special bound (Bound 1c for the norm-constrained case) was of the order of 10^6. Similar trends were observed in the VGG-26 network where our bounds were consistently tighter and outperformed the other bounds, providing considerably tighter generalization estimates.

Thus, our general bound (Bound 1a) was consistently smaller by a factor of 1000 compared to Bound 3 and by a factor of 100 compared to Bound 4, on all variants of the VGG network, which is a significant improvement. Moreover, our bound with the distance from initialization term (Bound 1b) was about half of that of the value of our general bound (Bound 1a) on all VGG network architectures. Most critically, our bounds were smaller than both, the size-independent results derived by Golowich et al. (2018) (Bound 5) and the recently obtained tighter bounds in Li et al. (2019) (Bound 2) by a factor of almost 10.

The general variant of our bound (1a) is tighter than the bounds 3 and 4 because the term inside the square root depends solely on the sum of the ratio of the $L_{2,1}$ norm and the spectral norm of the weights in each layer, which only grows linearly as $O(D)$ with an increasing number of layers. On the other hand, in Bound 3, the same term, being a cube of the sum of the ratio of the two weight norms, grows at a faster rate of $O(D^3)$, which is proportional to the cube of the depth of the network. Similarly, the term in Bound 4 has a factor D^2p multiplied to this sum of norm-ratios term, and thus grows with a rate $O(D) \times D^2p = O(D^3p)$, which is much larger than the rate at which our bound grows.

The results of these experiments confirm that our bounds are tighter than the current state-of-art norm-based generalization bounds for deep networks. Furthermore, our bound, under the per-unit weight constraints of 1, are even tighter and have better dependence on network dimensions, growing sub-logarithmically with depth while decreasing with increasing width of the network for a fixed depth. Furthermore, our bounds are based on the distance of network weights from their initialization instead of the norm of the weights, which is always smaller in magnitude according to the empirical observation. This empirical evidence of the fact that the network weight travels only a small distance from initialization during training allows us to derive tighter bounds.

Table 2: Comparison of different generalization bounds with our bounds on VGG-styled network with different number of layers. Our bound with the norm constraint on hidden units (Bound 1c) is the tightest of all bounds and is several order smaller than the previous bounds (Bounds 3 and 4).(Lecun et al., 1998)

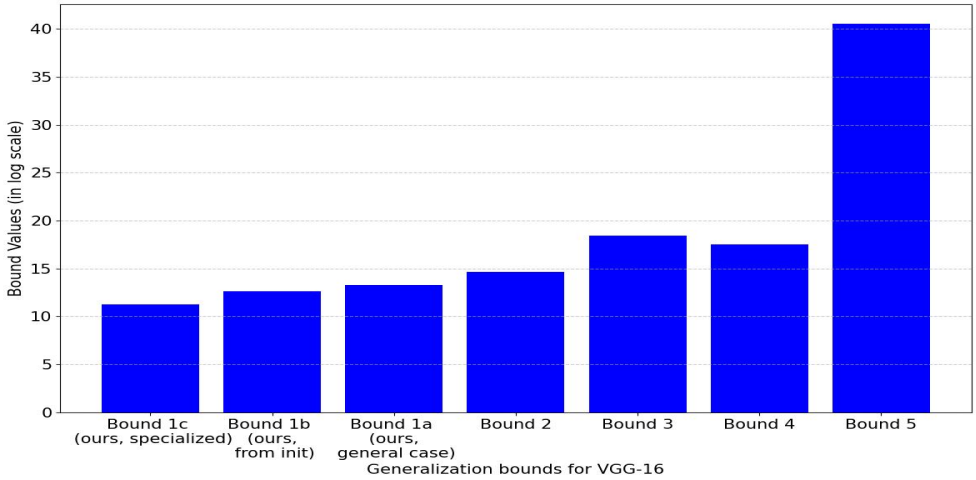
Model	Bound 1a	Bound 1b	Bound 1c	Bound 2	Bound 3	Bound 4	Bound 5
VGG-16	6×10^5	3×10^5	8×10^4	2.3×10^6	1×10^8	4×10^7	4.1×10^{17}
VGG-19	5.3×10^6	6.3×10^6	1.6×10^5	2.5×10^7	1.2×10^9	5.1×10^8	2.13×10^{22}
VGG-21	7.7×10^7	3.9×10^7	8.7×10^6	3.07×10^8	1.8×10^{10}	5.7×10^9	1.9×10^{24}
VGG-26	2.9×10^9	1.6×10^9	7.1×10^7	1.1×10^{10}	9.4×10^{11}	3.6×10^{11}	9.17×10^{30}

5 Discussion

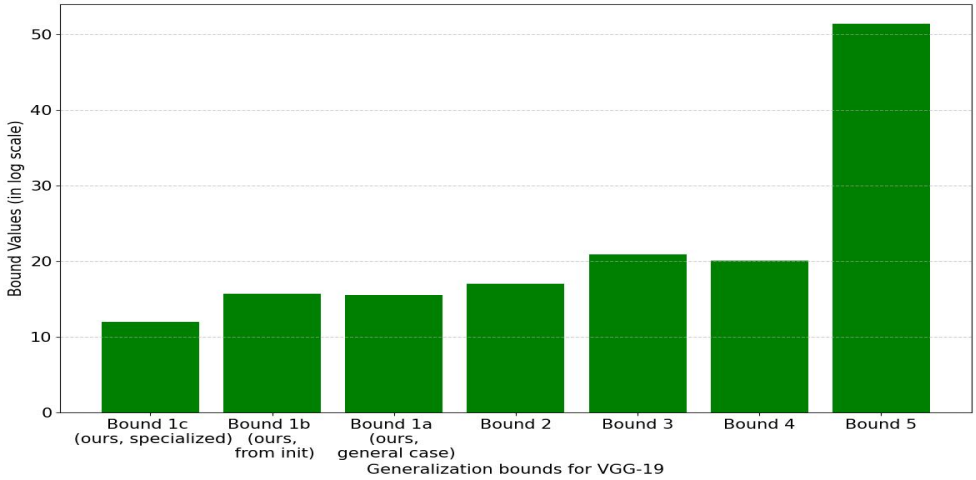
In this section, we critically examine and compare the tightness of our proposed generalization bounds against the related bounds discussed above based on the observed empirical results. As shown in Table 2, Bounds 3 and 4 were several orders of magnitude larger (specifically, 1000 and 100 times greater) than our proposed general bound (Bound 1a) for all the networks, and the difference was even more pronounced for our specialized bound in the norm-constrained case (Bound 1c), which was almost $10^3 \sim 10^4$ times smaller than these bounds. These observations align well with our theoretical predictions, according to which these bounds scale at a larger rate of $O(\sqrt{D^3pr})$ than our Bounds 1a and 1c that scale at rates of $(O(\sqrt{Dpr}))$ and $(O(\sqrt{log\ D}))$ respectively.

Surprisingly, our bounds demonstrated considerably tighter estimates, even when compared to the size-independent bounds (Bound 5) proposed by Golowich et al. (2018). We conjecture that this was due to the explicit dependence of this bound on the product of the Frobenius weight norm across all layers, which can grow exponentially faster than the product of spectral norm used in the other bounds (including ours). This is because the Frobenius norm is always larger than the spectral norm, since the Frobenius norm is proportional to the sum of the squares of all the singular values of the matrix, while the spectral norm is only dependent on the largest singular value. Moreover, for a layer d with p hidden units, the ratio of the Frobenius norm to the spectral norm satisfies the following:

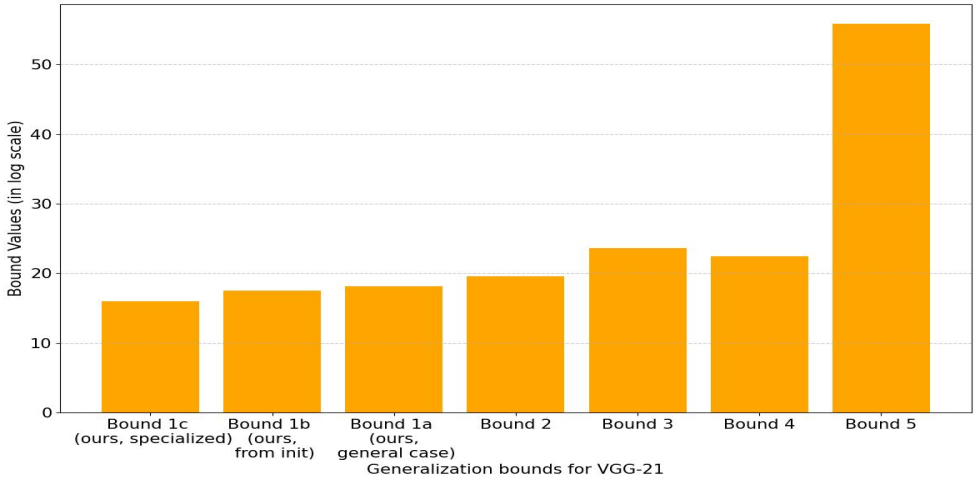
$$\frac{\|W_d\|_F}{\|W_d\|_2} = \sqrt{\frac{\sum_i \sigma_i^2}{\sigma_1^2}} \leq \sqrt{\frac{r\sigma_1^2}{\sigma_1^2}} = \sqrt{r} \leq \sqrt{p}$$



(a) Generalization bounds for VGG-16



(b) Generalization bound for VGG-19



(c) Generalization bound for VGG-21

Figure 1: Figure illustrating the values of the generalization bounds on a logarithmic scale for VGG-style networks of different depths (VGG-16 (subfigure a), VGG-19 (subfigure b), and VGG-21 (subfigure c)) (Bartlett et al., 2017) is much larger and loose for this network and is several orders larger than our bounds.

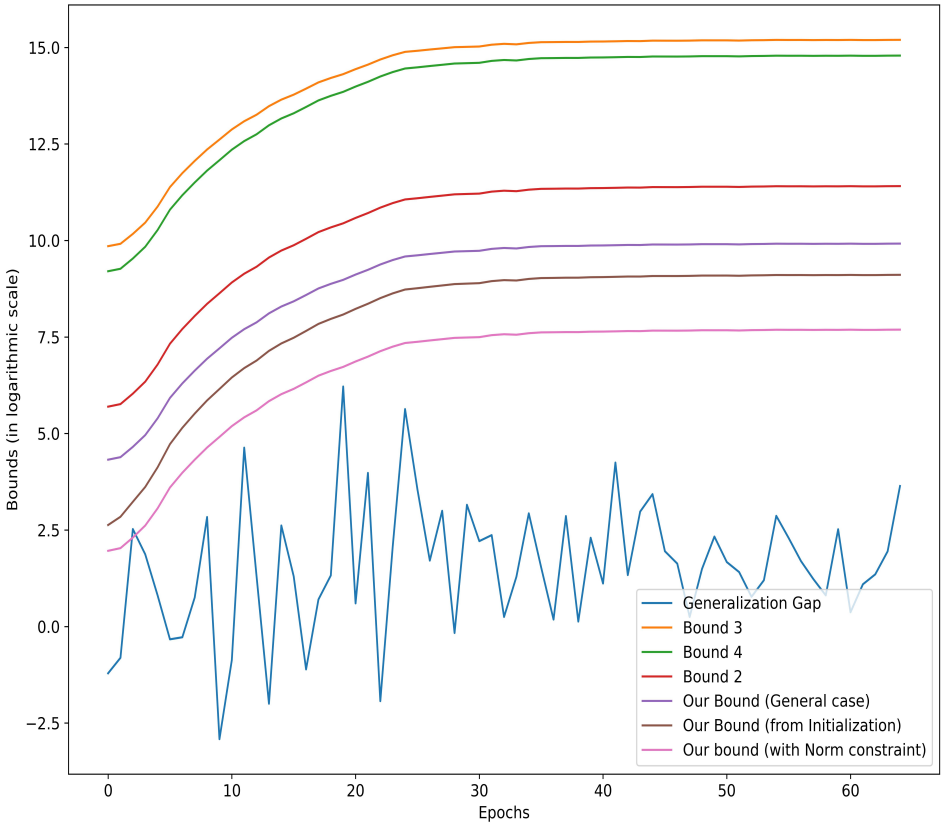


Figure 2: Figure illustrating the values of the different generalization bounds on a logarithmic scale as well as the generalization gap (difference between test and train loss) across different epochs for a VGG-21 model. Our proposed bound with the norm constraint on hidden units is the tightest bound and has the lowest value compared to the other bounds across all the epochs. On the other hand, Bound 3 (Bartlett et al., 2017) is much larger and loose for this network and is several orders larger than our bounds.

where r denotes the rank of the weight matrix which is always less than p. Thus, for a network with D layers, the ratio of the product of the two norms can grow exponentially according to the following inequality:

$$\frac{\prod_{d=1}^{D} \|W_d\|_F}{\prod_{d=1}^{D} \|W_d\|_2} \leq (\sqrt{p})^D \tag{17}$$

This fundamental difference in norm choice explains the faster growth rate of this bound with depth. Thus, this analysis leads us to posit that this bound's size independence relies upon the strong and often unrealistic assumption that the Frobenius norm of each layer be a very small quantity below 1, otherwise it risks an exponential increase with depth at a rate r^D, where $r>1$. This assumption is too restrictive and often may not hold in practical scenarios. In contrast, the tighter version of our bound (1c) is derived under a more relaxed and empirically plausible assumption: that the norm of each hidden unit weight is bounded by 1 (i.e., $\|W_{ik}\|_2 \leq 1$), which often holds well in practice. An upper bound on individual hidden units' norm still allows the Frobenius norm of each layer to grows up to $\sqrt{\sum_{j=1}^{p} \|W_{ij}\|^2} \leq \sqrt{p}$, leading to a more flexible and

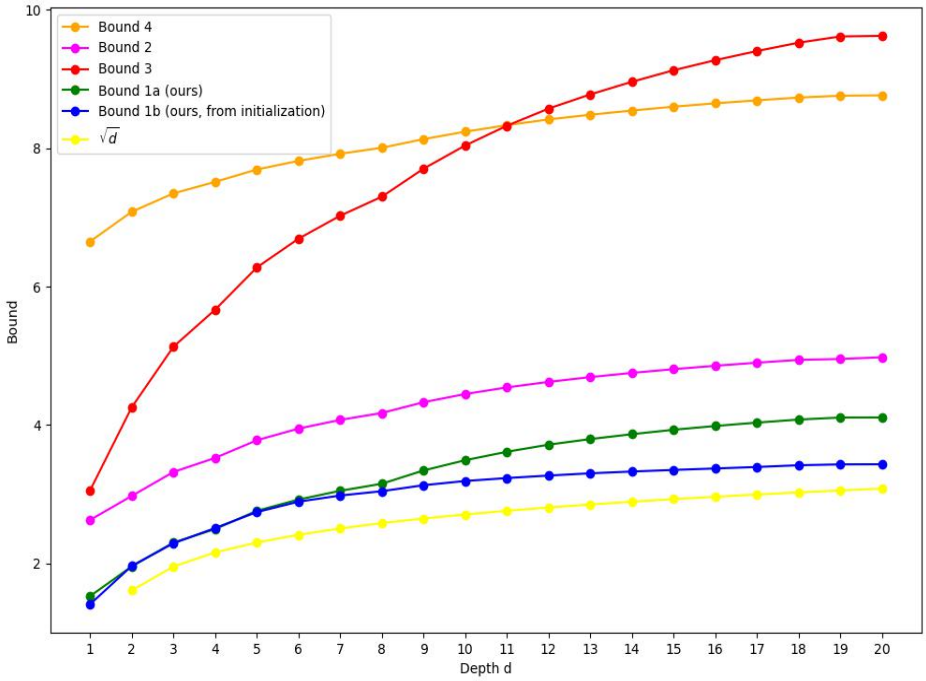


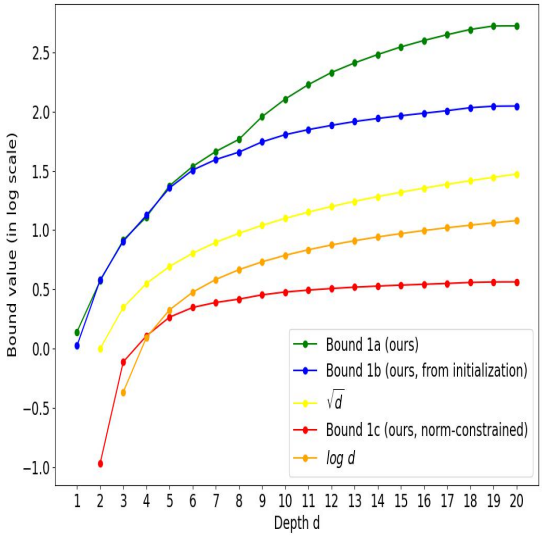
Figure 3: Figure illustrating the growth of different bounds with depth of a VGG-21 network. The horizontal axis in this figure corresponds to the depth, and the vertical axis represent the corresponding bound values in logarithmic scale.

realistic constraint for wide and overparameterized networks than the fixed and more stringent constraint that is placed on the Frobenius norm of each layer placed in Bound 5 ((Golowich et al., 2018)).

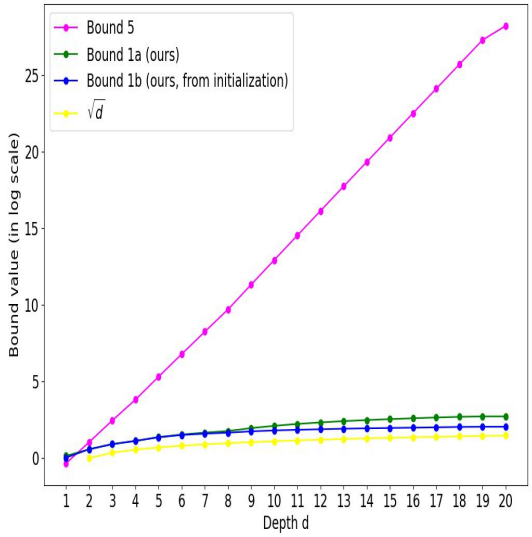
Furthermore, we observed that incorporating the distance of the weights from their initialization into the bounds always leads to tighter bounds compared to those that simply rely on the weight norm. This can be confirmed from Table 2, where we observe that the Bound 1b, which includes the distance from initialization term, was almost half of our bound in the general case (Bound 1a) for networks of different depths.

More critically, our empirical analysis reveals that our bounds were much tighter than even the recently proposed bounds by Li et al. (2019) (Bound 2) that used a term that explicitly depends on the network dimensions ($\sqrt{Dpr}$), whereas our bounds are mostly dependent on the network norms. As observed in Table 2, our Bound 1a was almost 10 times smaller than this bound, verifying its tighter estimates empirically. This demonstrates that our general bound (Bound 1a) increases at a rate that is much slower than the square root of the number of network parameters. Furthermore, our specialized bounds for the norm-constrained case (Bound 1c) are almost 100 times smaller than Bound 2 for networks of all sizes, highlighting the much tighter estimates provided by our bounds.

The results of these experiments suggest that although the works in Golowich et al. (2018) and Li et al. (2019) (Bounds 5 and 2) attempted to provide bounds that were almost size-independent or scaled more favorably with network dimension, our experimental observations suggest that their bounds are either still loose or are predicated on the assumptions of a constrained Frobenius norm in each layer, which are often larger compared to the spectral norm or the $L_{2,1}$ norm used in our bounds. On the other hand, our bounds not only provide more favorable scaling, but also provide much tighter estimates in the norm-constrained case. Furthermore, when deriving our logarithmic bounds, we impose a more relaxed constraint on the



(a) Comparison of growth rate with depth of our norm-constrained bound (1c) with our other bounds (1a, 1b), the square root of depth, and the log of the depth for a VGG-21 network.



(b) Demonstration of the explosive growth rate of the Bound 5 with depth when compared to our bounds.

Figure 4: In the subfigure (a), we demonstrate the tight scaling behavior of our norm-constrained bound (1c) with respect to our other bounds across different layers of depths d of a VGG-21 network. From the figure, we can see that our bound 1c grows at a smaller rate than $O(log\ d)$ and much slower than our other bounds as well as $O(\sqrt{d})$. In subfigure (b), we can observe that the Bound 5 (Golowich et al., 2018), which was purported to be size independent, grew at a much faster rate than our other bounds demonstrating its poor scaling.

weight norms of individual hidden units, which is significantly less restrictive and more realistic than the stringent constraint on the norm of the entire layer imposed in (Golowich et al., 2018) to derive their bounds.

6 Conclusion

In this work, a novel nonvacuous generalization bound demonstrating better dependence on network dimensions (width as well as depth) was introduced.

To provide a motivation for our bound, we began by outlining the shortcomings of traditional complexity bounds, such as those based on VC dimension and the principle of uniform convergence, which often yield vacuous bounds for deep over-parameterized networks. We then proceeded to provide an overview of prior nonvacuous bounds that attempted to overcome these shortcomings. These bounds were functions of the network norms and their margin. We discussed their strengths while also underscoring a significant limitation: their poor scaling with network dimensions either in the form $O(\frac{\sqrt{D^3pr}}{\sqrt{m}})$ or $O(r^D)$. We then introduced our bounds that provided significantly tighter estimates with better dependence on network dimensions (scaling with a rate of $O(\sqrt{Dpr})$). Moreover, under specific justifiable assumptions regarding the bounded norm of the network weights, our bounds achieved a remarkable sub-logarithmic scaling of $(O(\frac{\sqrt{log\ D}}{\sqrt{m}})$ with depth.

The theoretical rigor of our work is supported by extensive mathematical proofs of our bounds for both the general and the special bounded norm case. Empirical validation on VGG networks of different depths

demonstrated the significantly tighter estimates and better depth scaling provided by our bounds compared to the competitive nonvacuous bounds.

Our rigorous theoretical framework, combined with compelling experimental evidence, established the superiority of our bounds compared to the previous ones and also demonstrates the tighter depth as well as width dependence of our bounds compared to the previous work. Thus, our work provides novel theoretical insights into the remarkable generalization capabilities of very deep networks, especially in the over-parameterized regime. Our findings contribute to a deeper understanding of the reason behind the superior generalization performance of deep neural networks, despite their immense complexity.

7 Future directions

Our current work provides a robust foundation for a deeper understanding of generalization in deep networks, yet it also opens several avenues for future research to derive even tighter bounds that may provide a more precise estimate of network generalization.

One critical direction is the incorporation of the role of implicit regularization provided by Stochastic Gradient Descent (SGD). Although our derived bounds exclude the effect of this crucial component, many works have demonstrated that SGD indeed plays a vital role in biasing the trained network towards lower complexity solutions (Smith et al., 2021; Arora et al., 2019). It would be an interesting direction to analyse the effects of this optimization algorithm in greater detail and to include its impact on the generalization bound for the deep networks.

Furthermore, as shown in (Zhang et al., 2021) data plays an important role in determining how well the trained network will generalize to unseen data. As (Zhang et al., 2021) has shown that deep networks have sufficient capacity to fit even random data, but a network trained on such data will obviously generalize very poorly, the data distribution thus plays a significant role in determining the generalization property of the deep networks. It would be thus an important direction to explore the impact of the data distribution on the generalization bounds of the neural networks. Thus, in future works, we would like to extend our bounds to address these challenges and incorporate the role of the training algorithm and the data distribution in the derivation of more tighter and improved generalization bounds.

References

Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks, 2017. URL `https://arxiv.org/abs/1710.03667`.

Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. URL `https://arxiv.org/abs/1803.08375`.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf`.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017.

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 and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, March 2003. ISSN 1532-4435.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on*

Artificial Intelligence and Statistics, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

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

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond, 2019. URL https://arxiv.org/abs/1806.05159.

David McAllester. Simplified pac-bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth (eds.), *Learning Theory and Kernel Machines*, pp. 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

Vaishnavh Nagarajan and J. Zico Kolter. Generalization in deep networks: The role of distance from initialization, 2019. URL https://arxiv.org/abs/1901.01672.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning, 2021. URL https://arxiv.org/abs/1902.04742.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks, 2018. URL https://arxiv.org/abs/1707.09564.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent, 2021. URL https://arxiv.org/abs/2101.12176.

V. N. Vapnik and A. Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pp. 11–30. Springer International Publishing, Cham, 2015. ISBN 978-3-319-21852-6. doi: 10.1007/978-3-319-21852-6_3. URL https://doi.org/10.1007/978-3-319-21852-6_3.

Roman Vershynin. High-dimensional probability. 2018. URL https://api.semanticscholar.org/CorpusID:196128750.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

A Appendix

A.1 Additional Lemmas

In this section of the appendix, we will derive some additional lemmas which will be useful in deriving our generalization bound in the special norm-constrained case.

Lemma A1 *Given a neural network with D layers with a maximum of p hidden units in any layer. The covering number of the i_{th} layer with a rank r of its weight matrix is given by*

$$\Big(\frac{2b_i L_{W_i}}{\epsilon_i^2}\Big)^{2pr} \tag{18}$$

where L_{w_i} is the Lipschitz constant of the i_{th} layer, b_i bounds the sum of the distance of the hidden unit weights of layer i from their initialization ($\|W_i - W_{i_0}\|_{2,1} < b_i$), and ϵ_i denotes the radius of the cover element used in the i_{th} layer.

Furthermore, we will use the following lemma to bound the Lipschitz constant (L_{w_i}) of the i_{th} layer of the network:

Lemma A2 *Given a neural network $f_W(.)$ whose Jacobian with respect to all layers excluding the i_{th} layer is given by $B^{jac}_{\setminus i}$ and the maximum input norm is bounded by $\max_x \|x\|_2 < R$ then the Lipschitz constant of the margin loss function $g_\gamma(W, x)$ with respect to the i_{th} layer weights w_i can be upper-bounded as*

$$L_{w_i} < \frac{B^{jac}_{\setminus i} R}{\gamma} \tag{19}$$

Lemma A3 *(Bartlett et al., 2017) Assume $(\epsilon_1, \epsilon_2, ..., \epsilon_L)$ are given along with a set of matrices $W_1, W_2, .., W_i, .., W_L$ that lie within $\mathcal{B}_1 \times \mathcal{B}_2 \times .. \times \mathcal{B}_L$ where $\mathcal{B}_i$ are classes in $\mathbb{R}^p$ with variable p. and $\|W_i\| < c_i$. Let us define $\delta = \sum_{j \leq L} \rho_i c_i \prod_{l=j+1}^{L} \rho_l c_l$, then the hypothesis class representing the neural network outputs ($\mathcal{F}_\mathcal{L} = \{Z_\mathcal{W} : \mathcal{W} \in \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3 \times .. \times \mathcal{B}_L\}$) has the covering bound.*

$$\mathcal{N}(\mathcal{F}_\mathcal{L}, \delta, |.|_L) = \prod_{i=1}^{L} \sup_{W_1, W_2, .., W_{i-1}} \mathcal{N}\Big(\{W_i Z_{i, W_1, W_2 .., W_{i-1}}(.) : W_i \in B_i\}, \epsilon_i, \|.\|\Big) \tag{20}$$

Proof. To prove this theorem, we will break down the task of calculating the covering number for the i_{th} layer into a product of the covering number for the previous layer $(i-1)$ and the cover for the weights of the i_{th} layer. We will proceed with the proof inductively.

- Firstly, let's consider the covering bound for the first layer defined by the class $\mathcal{F}_1 = \{W_1 X : W_1 \in \mathcal{B}_\infty\}$ and denote it by $N_1 = \mathcal{N}(\mathcal{F}_1, \epsilon_1, |.|_2)$
- Assuming that the lemma holds for all layers upto i, for the function class in layer $i+1$: $\mathcal{F}_{i+1} = \{W_i Z_{W_1, W_2 .., W_{i-1}} : W_i \in \mathcal{B}_i\}$, construct a cover for the layer i weights individually defined as

$$|\mathcal{H}_{i+1}| \leq \sup_{(W_1, W_2, W_3, .., W_i)} \mathcal{N}(\{W_{i+1} Z_{W_1, W_2, .., W_i} : W_{i+1} \in \mathcal{B}_{i+1}\}) := N_{i+1} \tag{21}$$

  Now, using above, we can define the cover for the $(i+1)_{th}$ layer as follows

$$\mathcal{F}_{i+1} = \bigcup_{F \in \mathcal{F}_i} \mathcal{H}_{i+1} \tag{22}$$

$$|\mathcal{F}_{i+1}| \leq |\mathcal{F}_i| * |\mathcal{H}_{i+1}| \leq |\mathcal{F}_i| * N_{i+1} \leq N_{i+1} N_i |\mathcal{F}_{i-1}| \leq \prod_{j=1}^{i+1} N_j \tag{23}$$

Now, define functions $\mathcal{F} = \{\sigma(Z) : Z \in \mathcal{F}_L\}$, where $\sigma(.)$ is an activation function with Lipschitz constant c_L. Now, to demonstrate the existence of a cover for this set of functions, let us define a set of reference weights, which in our case are the weights set at initialization $(W_{1_0}, W_{2_0}, .., W_{L_0}) \in (\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3 \times ... \times \mathcal{B}_L)$

Using these, we can inductively define the bound on each layer using the following cases.

- Consider the base case $|\mathcal{H}_0| = \|W_0 - W_{0_0}\|\|X\|$
- Next choose the cover elements $\widehat{F}_i$ that satisfy $\|W_i\widehat{F_{i-1}} - \widehat{F}_i\| \leq \epsilon_i$ for each layer. Using this, we derive the bound

$$|F_i - \widehat{F}_i| < \sum_{j=1}^{i} \epsilon_j \rho_j \prod_{k=j+1}^{i} \rho_k c_k \tag{24}$$

  where F_i is the output of i_th neural network layer of the trained model (corresponding to weight W_i) and F_{i_0} are the outputs from the same layer at initialization. We prove this inductively below.

For the base case, we know $|F_0 - \widehat{F_0}| \leq \|W_0 - W_{0_0}\|\|X\| \leq \epsilon_0$. For the inductive step

$$\begin{aligned}
|F_{i+1} - \widehat{F_{i+1}}| &\leq |F_{i+1} - \rho_i W_i \widehat{F}_i + \rho_i W_i \widehat{F}_i - \widehat{F_{i+1}}| \\
&\leq \rho_i |W_i F_i - W_i \widehat{F}_i| + \rho |W_i \widehat{F}_i - \widehat{F_{i+1}}| \\
&\leq \rho_i |W_i||F_i - \widehat{F}_i| + \rho_i \epsilon_i \\
&\overset{(i)}{\leq} \rho_{i+1} c_{i+1} \Big(\sum_{j=1}^{i-1} \epsilon_j \rho_j \prod_{k=j+1}^{i-1} \rho_k c_k \Big) + \epsilon_{i+1} \rho_{i+1}
\end{aligned}$$

which, after simplification, gives back equation (24) for the $(i+1)_{th}$ layer. Here, the inequality marked (i) is derived by substituting (24) by induction.

A.2 Proofs

We now provide the proofs of the important lemmas and theorems stated earlier in the paper. Before describing their proofs, we introduce an additional lemma which will be useful in the analysis of covering numbers.

A.2.1 Proof of our initialization-based spectral covering bound (Lemma 1)

We now derive the proof of our covering bound described in Lemma 1. First, let us set up some notation and assume $\|W_i\|_2 \leq s_i$, $\|W_i - W_{i_0}\|_{2,1} \leq b_i$ and define the per-layer cover resolutions $(\epsilon_1, \epsilon_2, .., \epsilon_D)$ as follows:

$$\epsilon_i = \frac{\epsilon}{\rho_i \prod_{j>i} \rho_j s_j} \tag{25}$$

where ρ_i are the Lipschitz constant of the i_{th} layer.

Now, we use Lemma A1 to break down the covering number for the entire network into its individual layers and then aggregate them to the full network covering bound. Note that we use the matrix at initialization to describe the set to which the weights at each layer belongs ($\mathcal{B}_i = \{W_i : |W_i - W_{i_0}|_{2,1} < b_i\}$). We can now use the lemma A1 with these assumptions to prove our result

$$ln\ \mathcal{N}(\mathcal{F}_D, \epsilon, \|.\|) \leq \sum_{i=1}^{D} \underset{(W_0, W_1, .., W_{i-1})}{sup} ln\ \mathcal{N}(\{W_i Z_{W_1, W_2, .., W_{i-1}}(X) : W_i \in \mathcal{B}_i\}, \epsilon_i, \|.\|) \tag{26}$$

$$\sum_{i=1}^{D} \underset{(W_0, W_1, .., W_{i-1})}{sup} ln\ \mathcal{N}(\{W_i Z_{W_1, W_2, .., W_{i-1}}(X) : \|W_i - W_{i_0}\|_{2,1} \leq b_i, \|W_i\|_2 \leq s_i\}, \epsilon_i, \|.\|) \tag{27}$$

$$\overset{(i)}{\leq} \sum_{i=1}^{D} \underset{(W_0, W_1, W_2, .., W_{i-1})}{sup} \frac{b_i^2 \|Z_{W_1, W_2, ..., W_{i-1}}(X)\|_2^2}{\epsilon_i^2} ln\ (2W^2) \tag{28}$$

where the bound (28) is derived from (27) by using Lemma 3.2 from (Bartlett et al., 2017). In order to simplify this equation, we expand $\|Z_{W_1, W_2, ..., W_D}(X)\|_2^2 \leq \rho_L \|W_D Z_{W_1, W_2, .. W_{D-1}}(X)\|_2^2 \leq$

$\rho_D \|W_D\|_2^2 \|Z_{W_1,W_2,..W_{D-1}}(X)\|_2^2$. Proceeding inductively, this gives.

$$\underset{j}{max}\ \|Z_{W_1,W_2,...,W_D}(X_j)\|_2^2 \leq \|X\|_2 \prod_{i=1}^{D} \rho_i \|W_i\|_\sigma \tag{29}$$

Substituting this result in the inequality (28), we get

$$\begin{aligned} ln(\mathcal{F}_\mathcal{D}, \epsilon, \|.\|) &\leq \sum_{i=1}^{D} \underset{(W_1,W_2,..,W_D)}{sup} \frac{b_i^2 \|X\|_2^2 \prod_{j<i} \rho_i^2 \|W_i\|_\sigma^2}{\epsilon_i^2} \\ &\leq \sum_{i=1}^{D} \frac{b_i^2 \|X\|_2^2 \prod_{j<i} \rho_j^2 s_j^2}{\epsilon_i^2} \\ &\overset{(ii)}{\leq} \sum_{i=1}^{D} \frac{b_i^2 \|X\|_2^2 \prod_{j=1}^{D} \rho_j^2 s_j^2}{\epsilon^2} (\frac{1}{s_i^2}) \\ &\leq \sum_{i=1}^{D} \frac{\|X\|_2^2 \prod_{j=1}^{D} \rho_j^2 s_j^2}{\epsilon^2} \Big(\frac{b_i}{s_i}\Big)^2 \\ &\leq \frac{B^2 \prod_{j=1}^{D} \rho_j^2 s_j^2}{\epsilon^2} \sum_{i=1}^{D} \Big(\frac{b_i}{s_i}\Big)^2 \end{aligned}$$

where we obtain the inqeuality $\overset{(ii)}{\leq}$ by substituting ϵ_i from (25) in the expression. This completes the given proof. This inequality can thus be simplified as

$$ln(\mathcal{F}_D, \epsilon, \|.\|) \leq \frac{B^2 \prod_{j=1}^{D} \rho_i^2 \|W_i\|_\sigma^2}{\epsilon^2} \sum_{i=1}^{D} \Big(\frac{\|W_i - W_{i_0}\|_{2,1}}{\|W_i\|_\sigma}\Big)^2 \tag{30}$$

A.2.2 Proof of covering bound for single layer (Lemma A1)

Let us assume that the output of the i_{th} layer of the neural network with at most p hidden units is given by $f^i(w, x)$, where $f^i(.)$ belongs to the function space $f_i \in \mathcal{F}$. Then, the metric in this function space is given as

$$L_g = \underset{x}{sup}\ \|f_1^i(x) - f_2^i(x)\|_2 \leq \underset{x}{sup}\ \|f(w_1^i, x) - f(w_2^i, x)\|_2 \leq L_{w_i} \|w_1^i - w_2^i\|_2 \tag{31}$$

with Lipschitz constant L_{w_i} and the weights in each layer are bounded as $\|W^i - W_0^i\|_{2,1} < b_i$. We can now bound the difference of function output in (31) for row-wise changes in weight vectors and sum it across all rows to get the following inequality:

$$L_g = \underset{x}{sup}\ \|f_1^i(x) - f_2^i(x)\|_2 \leq \underset{x}{sup} \|\sigma(W_1^i x) - \sigma(W_2^i x)\|_2 \leq \underset{x}{sup} \left\| \begin{matrix} \sigma'(W_{21}^i x)(W_{11}^i x - W_{21}^i x) \\ \sigma'(W_{22}^i x)(W_{12}^i x - W_{22}^i x) \\ \vdots \\ \sigma'(W_{2p}^i x)(W_{1p}^i x - W_{2p} x) \end{matrix} \right\|_2 \tag{32}$$

Assuming that the activation function σ has a Lipschitz constant ρ, then the derivative σ' is upper bounded by ρ and the above can be simplified as follows:

$$\sup_x \left\| \begin{matrix} \rho(W^i_{11}x - W^i_{21}x) \\ \rho(W^i_{12}x - W^i_{22}x) \\ \vdots \\ \rho(W^i_{1p}x - W^i_{2p}x) \end{matrix} \right\|_2 \overset{(i)}{\leq} \sup_x \sum_{j=1}^{p} \|\rho(W^i_{1j} - W^i_{2j})x\|_2 \overset{(ii)}{\leq} \sup_x \rho \sum_{j=1}^{p} \|W^i_{1j} - W^i_{2j}\|_2 \|x\|_2$$

$$\leq \rho \sum_{j=1}^{p} \|W^i_{1j} - W^i_{2j}\|_2 R \leq \rho R \sum_{j=1}^{p} \|W^i_{1j} - W^i_{2j}\|_2 \leq \rho R \|W^i_1 - W^i_2\|_{2,1} \quad (33)$$

where inequality (i) in the above expression is obtained by the triangle inequality, inequality (ii) is obtained by using the multiplicative inequality for the spectral norm $\|AB\|_2 \leq \|A\|_2\|B\|_2$, and $R = \sup_x \|x\|_2$. Assuming that the layer-wise Lipschitz constant is defined as $L_{w_i}{=}\rho R$, we get:

$$L_g \leq \rho R \|W^i_1 - W^i_2\|_{2,1} \leq \rho R \|W^i_1 - W^i_0 + W^i_0 - W^i_2\|_{2,1} \leq \rho R(\|W^i_1 - W^i_0\|_{2,1} + \|W^i_2 - W^i_0\|_{2,1}) \leq L_{w_i} * 2b_i \quad (34)$$

Now, since we know the maximum value that can be achieved by the layer-wise metric L_g, the problem of finding the covering number for the function space can be transferred to finding that of finding the covering number for the weight space $\mathcal{W}$ that satisfies the bound $\|W^i - W^i_0\|_{2,1} \leq b_i$.

Since we know that the ϵ-covering number of a set $A \in \mathbb{R}^p$ with diameter D_A is bounded by $\left(\frac{D_A}{\epsilon}\right)^p$ (Vershynin, 2018). For our weight space $\mathcal{W} \in \mathbb{R}^{p\times d}$ with a rank of r, the diameter is bounded as $D_A = \sup_{W^i_1, W^i_2} \|W^i_1 - W^i_2\| = \sup_{W^i_2, W^i_2} \|W^i_2 - W^i_0 + W^i_0 - W^i_2\| \overset{(i)}{\leq} \sup_{W^i_1, W^i_2} \|W^i_1 - W^i_0\|_{2,1} + \|W^i_2 - W^i_0\|_{2,1} \leq 2 * b_i$, where (i) is obtained by applying the triangle inequality. Hence, the upper bound on the ϵ-covering number over the weight space $\mathcal{W}$ in layer i is given by $\left(\frac{2b_i}{\epsilon}\right)^{pd}$ or $\left(\frac{2b_i}{\epsilon}\right)^{D_{eff}}$ where D_{eff} is the effective dimensionality.

For a weight matrix with rank r, dimension d and p hidden units, the effective dimensionality is $D_{eff} = r(p + d - r)$. This is because any such matrix can be decomposed as $W = UV^T$, where U is a matrix of dimension $p \times r$ and V is a matrix of dimension $d \times r$, and both are of full rank with total effective dimensionality $= pr + dr$. Moreover, this decomposition is not unique, since we can multiply an invertible matrix A of rank $r \times r$ to obtain $W = (UA^{-1})(AV^T)$, all of which lead to the same decomposition. Thus, we will need to subtract r^2 from the effective dimensionality. Hence, the net effective dimensionality is D_{eff} $= pr + dr - r^2$.

Incorporating $D_{eff} = r(p+d-r)$, this bound can be written as $\left(\frac{2b_i}{\epsilon}\right)^{r(p+d-r)} \overset{(iii)}{\leq} \left(\frac{2b_i}{\epsilon}\right)^{2pr}$. (Here, inequality (iii) is derived using the facts that the input dimension to any layer $d \leq p$, the ratio $(2b_i/\epsilon > 1)$ assuming ϵ is small enough compared to b_i, and that the rank $r > 0$)).

Using the result for $\epsilon-$covering on the weight space, we can now translate this result to the function space $f \in \mathcal{F}$ as follows. Let us consider a set $C = \{W_1, W_2, ..W_N\}$ to be an $\epsilon - cover$ of the weight space. For each function $f^i(W, x)$ in the function space, we can consider a center function $f^i(W_c, x)$ with weights W_c belonging to the cover C such that $L_f(f(W; x), f(W_c; x)) \leq L_{w_i}\|W - W_c\|_{2,1} \leq L_{w_i}\epsilon$, which is the smallest distance of a function to the nearest central function. Now, to make a δ-cover in the function space, we need $L_{w_i}\epsilon \leq \delta$, or $\epsilon \leq \delta/L_{W_i}$. Thus, to get a δ cover in the function space, we will need a corresponding cover in weight space with $\epsilon = \delta/L_{W_i}$. Substituting this value of ϵ in our covering bound for the weight space, we get:

$$\mathcal{N}(\mathcal{F}, \delta, L_g) = \mathcal{N}(\mathcal{W}, \frac{\delta}{L_{w_i}}, \|.\|_{2,1}) \leq \left(\frac{2b_i}{\delta/L_{w_i}}\right)^{2pr} = \left(\frac{2L_{w_i}b_i}{\delta}\right)^{2pr} \quad (35)$$

Replacing the variable δ with ϵ in the above inequality, we finally complete the proof of the Lemma A1.

A.2.3 Proof of Lemma A2

Given the function representing the neural network with L layers as $f_{W_L}(f_{W_{L-1}}(..f_{W_d}..(f_{W_1}(x))..)..)$. We can derive the Lipschitz constant of the margin loss function $g_\gamma(f_W(x), y)$ with respect to i_{th} layer weights of the network by considering the change in the loss function for the different configurations of the weights in layer d:

$$
\begin{aligned}
&\sup_{W_{d1},W_{d2}} \|g_\gamma(f_{W_L}(..f_{W_{d1}}..(f_{W_1}(x))..)..), y) - g_\gamma(f_{W_L}(..f_{W_{d2}}..(f_{W_1}(x))..)..), y)\|_2 \overset{(a)}{\leq} \\
&\qquad \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|(f_{W_L}(..f_{W_{d1}}..(f_{W_1}(x))..)..) - (f_{W_L}(..f_{W_{d2}}..(f_{W_1}(x))..)..)\|_2 \\
&\qquad \sup_{W_{d1},W_{d2}} \frac{1}{\gamma} \leq \|J_L f_{W_{L-1}}(..f_{W_{d1}}..(f_{W_1}(x))..) - J_L f_{W_{L-1}}(..f_{W_{d2}}..(f_{W_1}(x))..)\|_2 \\
&\leq \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|J_L J_{L-1} J_{L-2}..f_{W_{d1}}(f_{W_{d-1}}(..f_{W_1}(x))..) - J_L J_{L-1} J_{L-2}..f_{W_{d2}}(f_{W_{d-1}}(..f_{W_1}(x))..)\|_2 \\
&\leq \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|J_L J_{L-1} J_{L-2}..J_{d+1}\|_2 \|f_{W_{d1}}(f_{W_{d-1}}(..f_{W_1}(x))..) - f_{W_{d2}}(f_{W_{d-1}}(..f_{W_1}(x))..)\|_2 \\
&\overset{(i)}{\leq} \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|J_L J_{L-1} J_{L-2}..J_{d+1}\|_2 \|f_{w_{d1}}(\rho_d\|W_{d-1}\|_2\rho_{d-1}\|W_{d-2}\|_2...\rho_1\|W_1\|X) \\
&\qquad - f_{W_{d2}}(\rho_d\|W_{d-1}\|_2\rho_{d-1}\|W_{d-2}\|_2...\rho_1\|W_1\|X)\|_2 \\
&\overset{(ii)}{\leq} \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|J_L J_{L-1} J_{L-2}..J_{d+1}\|_2 \|\rho_{d+1}W_{d1}(\rho_d\|W_{d-1}\|_2\rho_{d-1}\|W_{d-2}\|_2...\rho_1\|W_1\|) \\
&\qquad - \rho_{d+1}W_{d2}(\rho_d\|W_{d-1}\|_2\rho_{d-1}\|W_{d-2}\|_2...\rho_1\|W_1\|)\|_2\|X\|_2 \\
&\overset{(iii)}{\leq} \sup_{W_{d1},W_{d2}} \|J_L J_{L-1} J_{L-2}..J_{d+1}\|(\rho_{d+1}W_{d1}\|J_{d-1}J_{d-2}..J_1\| - \rho_{d+1}W_{d2}\|J_{d-1}J_{d-2}..J_1\|_2)\|\|X\|_2 \\
&\leq \sup_{W_{d1},W_{d2}} \frac{1}{\gamma}\|J_L J_{L-1} J_{L-2}..J_{d+1}\|(\rho_{d+1}W_{d1} - \rho_{d+1}W_{d2})\|J_{d-1}J_{d-2}..J_1\|\|X\|_2 \\
&\leq \sup_{W_{d1},W_{d2}} \frac{1}{\gamma} B^{jac}_{L:d+1} B^{jac}_{d-1:1}\|W_{d1} - W_{d2}\|_2\|X\|_2 = L_{W_d}\|W_{d1} - W_{d2}\|_2 \qquad (36)
\end{aligned}
$$

Here, the inequality (a) is derived using the definition of margin loss function defined in the Preliminaries section in the Introduction, (i) is obtained via the Cauchy-Scwartz inequality. The inequality (ii) is derived from (i) by factoring out the common term $\|X\|_2$ and using the fact that $\|f_{w_{d1}}(x) - f_{w_{d2}}(x)\|_2 \leq \|\rho_{d+1}(W_{d1} - W_{d2})X\|_2 \leq \rho_{d+1}\|W_{d1} - W_{d2}\|_2\|X\|_2$, where ρ_{d+1} is the Lipschitz constant for the d_{th} layer. Comparing the value of L_{w_d} in the above equation, we get the Lipschitz constant of the d_{th} layer weights as follows

$$
L_{w_d} = B^{jac}_{L:d+1} B^{jac}_{d-1:1} R/\gamma = \frac{B^{jac}_{/d} R}{\gamma} \qquad (37)
$$

where we substitute $\|X\|_2 = R$. This completes the proof of our lemma.

A.2.4 Proof of Theorem 3

To prove this theorem, we utilize the following upper bound relating the Rademacher complexity of the margin loss function, with respect to the network's weights, to its covering number:

$$\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_\mathcal{D})) = \underset{\beta>0}{inf}\left(\beta + \int_\beta^B \sqrt{log\ \mathcal{N}(\mathcal{G}_\gamma, \epsilon, L_2(S))}\right) \tag{38}$$

where $\mathcal{N}(\mathcal{G}_\gamma, \epsilon, L_2(S))$ denotes the ϵ-covering of the margin loss function class $\mathcal{G}_\gamma$ with respect to the metric $L_2(S)$, $\mathcal{F}_D$ denotes the function class represented by a D layer neural network, $B = \underset{f\in\mathcal{F},x}{sup}|g_\gamma(f,x)|$ and $g_\gamma(f_w(x), y) \in \mathcal{G}$ denotes the margin loss function for input x with corresponding label y, induced by the network with parameters w.

Now, given a network of depth D, we can break down the entire network covering number $\mathcal{N}(\mathcal{G}_L, \epsilon, L_2(S))$ into a product of the covering number of the individual layers as follows (Bartlett et al., 2017).

$$\mathcal{N}(\mathcal{G}_\gamma, \epsilon, L_2(S)) = \prod_{i=1}^{D} \underset{W_1,W_2,..W_{i-1}}{sup} \mathcal{N}(\{g_\gamma(W_i, x|W_1, W_2, ..W_{i-1}) : W_i \in \mathcal{W}\}, \epsilon_i, \|.\|) \tag{39}$$

where W_i denotes the weights of the layer i, $\mathcal{W}$ denotes the set of all possible values for the weights of different layers.

Now, we use the covering number bound for the individual layers as described in Lemma A1 and substitute it into the product of covering bounds in equation (39) to get the covering bound for the whole network:

$$\mathcal{N}(\mathcal{G}_L, \epsilon, L_2(S)) = log\ \prod_{i=1}^{D}\left(\frac{2b_iL_{w_i}}{\epsilon^2}\right)^{2pr} \tag{40}$$

Substituting it into equation (30), we obtain the Rademacher complexity for the full network:

$$\hat{\mathfrak{R}_n}(\mathcal{G}_\gamma(\mathcal{F}_D)) = inf_{\beta>0}\left(\beta + \frac{1}{\sqrt{m}}\int_\beta^B \sqrt{log\ \prod_{i=1}^{D}\left(\frac{2b_iL_{w_i}}{\epsilon^2}\right)^{2pr}} d\epsilon\right) \tag{41}$$

$$\leq inf_{\beta>0}\left(\beta + \frac{1}{\sqrt{m}}B\sqrt{log\ \prod_{i=1}^{D}\left(\frac{b_iL_{w_i}}{\beta^2}\right)^{2pr}} d\epsilon\right) \tag{42}$$

Substituting $\beta = 2\sqrt{Bpr\sqrt{2p}}$, we get

$$\hat{\mathfrak{R}_n}(\mathcal{G}_\gamma(\mathcal{F}_D)) = \sqrt{B}2^{5/4}p^{3/4}r^{1/2} + \frac{B}{\sqrt{m}}\sqrt{log\ \prod_{i}^{D}\left(\frac{2b_iL_{w_i}}{4Bpr\sqrt{2pr}}\right)^{2p}} d\epsilon \tag{43}$$

Moreover,

$$\begin{aligned} B = |\sup_{f\in\mathcal{F},x} g_\gamma(f,x)| &\leq \frac{\|\sup_{f\in\mathcal{F}} f(x)\|_2}{\gamma} \\ \leq \frac{\|f_{W_d}(f_{W_{d-1}}(..f_{W_1}(x)..)\|_2}{\gamma} &\leq \frac{\|J_{W_d}f_{W_{d-1}}(..f_{W_1}(x)..)\|_2}{\gamma} \\ \leq \frac{\|J_{W_d}J_{W_{d-1}}(..f_{W_1}(x)..)\|_2}{\gamma} &\leq \frac{\|J_{W_d}J_{W_{d-1}}J_{W_{d-2}}...J_{W_1}\|_2\|X\|_2}{\gamma} \\ &\leq \frac{B_{1:d}^{jac}R}{\gamma} \end{aligned}$$

Substituting the values of B and L_{w_i} (Lemma A2) into the log term in Equation (43), many of the terms are canceled and we obtain:

$$\hat{\mathfrak{R}}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) = 2\sqrt{Br}2^{1/4}(p)^{3/4} + \frac{B}{\sqrt{m}}\sqrt{log\ \prod_i^D \Big(\frac{b_i B_{\setminus i}^{jac} R}{2B_{1:d}^{jac} Rpr\sqrt{2p}}\Big)^{2pr} d\epsilon} \tag{44}$$

$$\implies \hat{\mathfrak{R}}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) = B\Big(\frac{2^{5/4}p^{3/4}r^{1/2}}{\sqrt{B}} + \frac{1}{\sqrt{m}}\sqrt{log\ \prod_i^D \Big(\frac{b_i}{\|W_i\|_2 2pr\sqrt{2p}}\Big)^{2pr} d\epsilon} \tag{45}$$

Where (45) is obtained from (44) by decomposing the Jacobian of the whole network B as $B_{1:d}^{jac} = B_{1:i-1}^{jac}\|W_i\|_2 B_{i+1:d}^{jac}$ and canceling common terms. Since, we assume that the L_2 norm of each hidden unit j is bounded by 1 ($\|W_{ij} - W_{ij_0}\|_2 < 1$) and the fact that the spectral norm is bounded by the frobenius norm (i.e., $\|W_i\|_2 \leq \|W_i\|_F$). Moreover, we know that

$$\|W_i\|_2 \leq \|W_i\|_F = \sqrt{\sum_{j=1}^p \sum_k (W_{ijk})^2}$$
$$= \sqrt{\sum_{j=1}^p \|W_{ij}\|_2^2} \overset{(i)}{\leq} \sqrt{\sum_{j=1}^p 1} \leq \sqrt{p} \approx O(\sqrt{p})$$

where inequality (i) follows by the assumption that the norm of each hidden unit weight is constrained by 1. Hence, $\|W_i\|_2 \leq \|W_i\|_F \leq O(\sqrt{p})$. Moreover, the b_i are bounded as

$$b_i = \sum_{j=1}^p \sqrt{\sum_k (W_{ijk} - W_{ijk_0})^2}$$
$$= \sum_{j=1}^p \|W_{ij} - W_{ij_0}\|_2 \leq \|W_{ij} + (-W_{ij_0})\|_2 \leq \|W_{ij}\|_2 + \|W_{ij_0}\|_2 \leq \sum_{j=1}^p 2 \leq 2p$$

and hence $b_i \leq O(p)$.

Hence, the individual terms inside the brackets in the *log* (i.e., $\frac{b_i}{\|W\|_2 2pr\sqrt{2p}}$) in the above equation would be of order $O(2p/(2pr \times \sqrt{p} \times \sqrt{2p})) \approx O(1/pr\sqrt{2})$ and decrease with increasing hidden units. Furthermore, this term would be less than 1 for any $p \geq 1$ and $r \geq 1$, which always holds true.

Now, since each term inside the product term in the *log* in equation (45) would be smaller than 1, we would be able to employ the inequality

$$iff \quad \forall_i\ x_i \leq 1$$
$$\implies \prod_{i=1}^D x_i < \sum_{i=1}^D x_i$$
$$\implies log(\prod_{i=1}^D x_i) < log(\sum_{i=1}^D x_i)$$

which is a form of Jensen's inequality. Hence, using this inequality in the above equation, we can convert the product into sums, and can simplify equation (40) as follows:

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{G}_\gamma(\mathcal{F}_D)) &= B\Big(\frac{2^{1/4}p^{3/4}r^{1/2}}{\sqrt{B}} + \frac{1}{\sqrt{m}}\sqrt{log\ \prod_{i=1}^{D}\Big(\frac{b_i}{\|W_i\|_2 2pr\sqrt{2p}}\Big)^{2pr}}\,d\epsilon \\
&\leq B\Big(\frac{2^{5/4}p^{3/4}r^{1/2}}{\sqrt{B}} + \frac{1}{\sqrt{m}}\sqrt{log\ \sum_{i=1}^{D}\Big(\frac{b_i}{\|W_i\|_2 2pr\sqrt{2p}}\Big)^{2pr}}\,d\epsilon \\
&\leq B\Big(\frac{2^{5/4}p^{3/4}r^{1/2}}{\sqrt{B}} + \frac{1}{\sqrt{m}}\sqrt{log\ \sum_{i=1}^{D}2pr\Big(\frac{b_i}{\|W_i\|_2 2pr\sqrt{2p}}\Big)}\,d\epsilon \\
&\leq B\Big(\frac{2^{5/4}p^{3/4}r^{1/2}}{\sqrt{B}} + \frac{1}{\sqrt{m}}\sqrt{log\ \sum_{i=1}^{D}\Big(\frac{b_i}{\|W_i\|_2 \sqrt{2p}}\Big)}\,d\epsilon
\end{aligned}$$

Substituting the value of $b_i = \|W_i - W_{i_0}\|_{2,1}$ and $B = \frac{B^{jac}_{1:D}R}{\gamma}$, the proof of the Theorem is completed.