A STRUCTURAL APPROACH TO PATH-NORM-BASED GENERALIZATION BOUNDS FOR NEURAL NETWORKS CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

While the path-norm has emerged as a theoretically grounded measure of neural network generalization, existing generalization bounds often rely on restrictive assumptions—such as the positive homogeneity of activation functions—and mostly target basic feedforward or residual architectures. In this paper, we develop an analytical framework that broadens the scope of path-norm theory to modern neural network architectures without relying on potentially loose covering-number estimates. First, by introducing the Absolute Rademacher Complexity, we remove the traditional positive homogeneity requirement and extend path-norm-based bounds to arbitrary Lipschitz-continuous activation functions. Second, we establish continuous-limit path-norm bounds for Neural Ordinary Differential Equations (Neural ODEs). By proving the uniform convergence of Rademacher complexities under Euler discretization, we characterize their generalization complexity through a differential comparison system. Finally, we decompose the self-attention mechanism and combine an element-wise matrix inequality with vector contraction to derive explicit, layer-wise bounds for Transformers. Our structural decomposition approach avoids the theoretical bottlenecks of traditional covering-number methods and provides architectural insights into how continuous-depth dynamics and attention operators contribute to model capacity.

1 INTRODUCTION

It is widely recognized that the scaling law is a pivotal phenomenon in neural network practice by Kaplan et al. (2020). According to classical machine learning perspectives Geman et al. (1992), a model with an excessively large number of parameters would be highly prone to overfitting; however, empirical evidence on double descent (Belkin et al. (2019)) suggests otherwise. A prominent research direction involves redefining model capacity: theoretical studies indicate that the upper bound of model capacity is governed not only by the number of parameters but also by the norm magnitude of the model(Liu et al. (2024)).

In our paper, we will extend the definition of path norm of neural network. According to the empirical findings by Jiang et al. (2020), the path norm serves as an effective indicator of generalization error. Path norms have been empirically demonstrated to correlate well with generalization, correctly capturing the effect of network depth. Furthermore, practical empirical studies Latorre et al. (2023) have attempted to regularize the path norm to enhance the generalization performance of deep learning models, a strategy that has been empirically proven successful. Mean while, traditional L2-norm even show negative relativity in certain cases(Jiang et al. (2020)).

However, existing research on path norms remains predominantly empirical and primarily focuses on simpler neural networks(Hashimoto et al. (2024),Golowich et al. (2018)), such as Multi-Layer Perceptrons (MLPs) and ResNets, lacking rigorous theoretical generalization bounds under more complex architectures. In this paper, we extend the analysis of path norms to two advanced network architectures: Neural Ordinary Differential Equations (Neural ODEs) and Transformers. By tailoredly defining the path norms for these advanced structures, we establish direct, path-norm-based generalization error bounds.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In terms of the analytical methodology, we propose a brand-new framework: we directly analyze the network's generalization error by decomposing its specific architecture, rather than indirectly estimating deep learning generalization errors through traditional covering number methods. This approach may provide deeper insights into our understanding of neural network structures.

To summarize, our main contributions are as follows:

- **Theoretical Extension to Advanced Architectures:** We extend the path norm analysis to sophisticated modern architectures, specifically Neural ODEs and Transformers, broadening the scope of path norm theory.
- **Direct and Tighter Generalization Bounds:** By tailoring path norm definitions to these complex structures, we derive direct generalization error bounds that bypass the limitations of traditional metrics.
- **Novel Deconstructive Methodology:** We propose a brand-new analytical approach based on structural decomposition rather than covering numbers, providing a deeper conceptual understanding of how network topologies affect generalization.

2 RELATED WORK

Generalization bound for neural networks

Authors	Rate	Type
Neyshabur et al. (2015)	$\frac{2^L \prod_{j=1}^L \|W_j\|_{2,2}}{\sqrt{n}}$	Norm-based
Neyshabur et al. (2018)	$\frac{L \max_j d_j \prod_{j=1}^L \|W_j\|}{\sqrt{n}} \left(\sum_{j=1}^L \frac{\|W_j\|_{2,2}^2}{\|W_j\|^2}\right)^{1/2}$	
Golowich et al. (2018)	$\left(\prod_{j=1}^L \|W_j\|_{2,2}\right) \min\left\{\frac{1}{n^{1/4}}, \sqrt{\frac{L}{n}}\right\}$	
Bartlett et al. (2017)	$\frac{\prod_{j=1}^L \|W_j\|}{\sqrt{n}} \left(\sum_{j=1}^L \frac{\|W_j^T - A_j^T\|_{2,1}^{2/3}}{\|W_j\|^{2/3}}\right)^{3/2}$	
Wei & Ma (2020) (2020)	$\frac{(\sum_{j=1}^L \kappa_j^{2/3} \min\{L^{1/2}\|W_j - A_j\|_{2,2}, \|W_j - B_j\|_{1,1}\}^{2/3})^{3/2}}{\sqrt{n}}$	
Ju et al. (2022)	$\frac{\sum_{j=1}^L \theta_j \|W_j - A_j\|_{2,2}}{\sqrt{n}}$	
Li et al. (2021)	$\|\mathbf{x}\| \prod_{j=1}^L \|W_j\| - 1\| + \gamma_{\mathbf{x}} + \sqrt{\frac{c_{\mathbf{X}}}{n}}$	

Table 1: Generalization bound of multiple-layer neural network

κ_j and θ_j are determined by the Jacobian and Hessian of the network f with respect to the jth layer and W_j, respectively. $\tilde{r}_j$ and $\tilde{d}_j$ are the rank and dimension of the jth weight matrices for the compressed network.

While classical generalization bounds heavily rely on spectral or L_2-norms (e.g., Neyshabur et al. (2015); Bartlett et al. (2017)), extensive empirical studies—such as Jiang et al. (2020)—have demonstrated that L_2-based metrics often exhibit weak or even negative correlation with true generalization error. In contrast, the path-norm has emerged as a robust and highly correlated predictor of model performance. However, historical attempts to establish path-norm-based generalization bounds are plagued by severe architectural bottlenecks. Existing proofs (Golowich et al. (2018)) fundamentally depend on architecture-specific mathematical tricks (such as the positive homogeneity of ReLU) tailored for simple Multi-Layer Perceptrons (MLPs). Consequently, these traditional techniques fail when confronted with the complex topologies of modern architectures, such as the skip connections in ResNets, the continuous dynamics of Neural ODEs, or the intricate self-attention mechanisms in Transformers.

Generalization bound for Neural ODE

Recent theoretical analyses of Neural Ordinary Differential Equations (ODEs) have struggled to balance optimal sample complexity with high-dimensional scalability.

- Verma & Kumar (2025): By analyzing the bounded variation properties of non-linear Neural ODE solutions, this work established a generalization bound that achieves the standard $\mathcal{O}(1/\sqrt{n})$ sample convergence. However, their reliance on covering numbers introduces a severe dimensional penalty, resulting in a bound that scales as $\mathcal{O}(d^{3/2}/\sqrt{n})$ where d is input dimension.
- Marion (2023) Prior to this, bounds for continuous-in-time parameterized ODEs and deep residual networks were derived using Lipschitz continuity assumptions. Dealing with the infinite-dimensional parameter space forced a mathematical compromise, yielding a sub-optimal convergence rate that scales as $\mathcal{O}(d^2/n^{1/4})$.

In contrast to these approaches, our work establishes a $\mathcal{O}(1/\sqrt{n})$ generalization bound that only depends on norm of parameters. We successfully recover the optimal sample complexity while completely decoupling the error bound from the state-space dimension d, overcoming the inherent mathematical bottlenecks present in prior bounded-variation and covering-number techniques.

Generalization bound for transformer

While Transformer generalization has been explored via PAC-Bayes (Lotfi et al. (2024)) and algorithmic stability (Li et al. (2023); Yao et al. (2025)), explicit complexity bounds remain limited. Existing norm-based approaches often rely on coarse global spectral approximations (Edelman et al. (2022); Trauger & Tewari (2024)). Recently, Mwigo & Dasgupta (2026) derived a Rademacher complexity bound, but their analysis is strictly confined to shallow (single-layer) architectures under restrictive lazy-training regimes.

3 PATH-NORM BASED RADEMACHER COMPLEXITY

3.1 PRELIMINARIES

Recall definition of Rademacher complexity: Let $\mathcal{F}$ be a function class, $x_1, , ..., x_n$ are data points and $\varepsilon_1, ..., \varepsilon_n$ be independent and identically distributed random variables uniformly distributed in $\{-1, 1\}$. Then Rademacher complexity of $\mathcal{F}$ with respect to $\{x_1, ..., x_n\}$ is defined as following:

$$Rad(\mathcal{F}) = E_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)$$

The pivotal role of Rademacher complexity lies in its ability to provide a distribution-free upper bound on the generalization error. Let $\mathcal{D}$ be the underlying data distribution, and $S = \{x_1, \ldots, x_n\}$ be a training dataset sampled i.i.d. from $\mathcal{D}$. For a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, we define the empirical risk and the expected risk of a hypothesis $f \in \mathcal{F}$ as follows:

$$R_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i), \quad R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)].$$

By applying McDiarmid's inequality, we can bound the uniform deviation between the expected and empirical risks. Specifically, we have the following fundamental theorem linking Rademacher complexity and generalization:

Theorem 1 (Generalization Bound via Rademacher Complexity, Bartlett & Mendelson (2002))**.** *Let $\mathcal{G} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$ be the loss function class associated with $\mathcal{F}$, where the loss function takes values in $[0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the sample S of size n, the following inequality holds for all $f \in \mathcal{F}$:*

$$R(f) - R_S(f) \leq 2Rad(\mathcal{G}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

This standard framework serves as the foundation for our generalization analysis. In the subsequent sections, we will show how the unique structural properties of Neural ODEs and Transformers allow us to directly decompose and control Rad($\mathcal{F}$) via our newly proposed path norms, without resorting to covering number bounds.

3.2 PATH NORM

For a d-layer MLP:$\mathbb{R}^n \to \mathbb{R}$ with no bias term: $NN(x) = f_d(g(f_{d-1}(\ldots\ldots g(f_0(\mathbf{x})))))$, where g is activation function satisfying 1-Lipschitz, $g(0) = 0$, $f_k(\mathbf{x}) = W_k\mathbf{x}$, its path norm is defined as: $\prod_{i=n}^{1} |W_i|\mathbf{1}^T$, where $|W_i|$ refers to we taking element-wise absolute value, and $\mathbf{1}$ refers to row vector that every component is 1. A natural generalization of path norm to MLP with bias term is defined as following: For a d-layer MLP:$\mathbb{R}^n \to \mathbb{R}$ with bias term: $NN(x) = f_d(g(f_{d-1}(\cdots g(f_0(\mathbf{x})))))$, where g is activation function satisfying 1-Lipschitz, $g(0) = 0$, $f_k(\mathbf{x}) = W_k\mathbf{x} + b_k$, its path norm is defined as:

$$F_d(\cdots F_1(\mathbf{1}^T)\cdots)$$

Where $F_i(x) = |W_i|x + |b_i|$. $|W_i|, |b_i|$ refers to we taking element-wise absolute value, and $\mathbf{1}$ refers to row vector that every component is 1.

3.3 LEMMA

Before establishing our major result, Before proving our main result, we provide the following lemmas as preparation. These lemmas play a central role in our proofs and computations.

Lemma 1. *Let $\mathcal{F}_{\alpha,\theta}$ be parametrized function class depending on α and θ; $\mathcal{G}_{\beta,\theta}$ be parametrized function class depending on β, θ. α, β, θ are restricted in compact domains. θ are common parameters shared by $\mathcal{F}$ and $\mathcal{G}$. Then*

$$Rad(\mathcal{F}_{\alpha,\theta} + \mathcal{G}_{\beta,\theta}) \leq Rad(\mathcal{F}_{\alpha,\theta}) + Rad(\mathcal{G}_{\beta,\theta})$$

Lemma 2. *Let $\mathcal{F}$ be function class where $0 \in \mathcal{F}$. L is some fixed real number and α is a real variable satisfying $|\alpha| \leq \frac{L}{2}$. Then*

$$Rad(\alpha\mathcal{F}) \leq LRad(\mathcal{F})$$

Lemma 3 (Talagrand's Contraction Lemma, Ledoux & Talagrand (1991))**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions. Then, for any function class $\mathcal{F}$ of real-valued functions, the following inequality holds:*

$$\mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(f(x_i)) \right] \leq L \cdot \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right] \tag{1}$$

Specifically, if g is a 1-Lipschitz activation function (such as ReLU) satisfying $g(0) = 0$, we have:

$$Rad(g \circ \mathcal{F}) \leq Rad(\mathcal{F}) \tag{2}$$

In addition, we use another concept to measure the complexity of function families, namely the absolute Rademacher complexity. For dataset $S = x_1, \cdots, x_n$, it is defined as follows:

$$Rad^{abs}(\mathcal{F}) = \frac{1}{n} E_\varepsilon \sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i f(x_i)|$$

By direct checking definition, we have following equation:$Rad^{abs}(\mathcal{F}) = Rad(\mathcal{F} \cup (-\mathcal{F}))$. Moreover, we claim following lemma on absolute Rademacher complexity:

Lemma 4. *Let $\mathcal{F}_{\alpha,\theta}$ be parametrized function class depending on α and θ; $\mathcal{G}_{\beta,\theta}$ be parametrized function class depending on β, θ. α, β, θ are restricted in compact domains. θ are common parameters shared by $\mathcal{F}$ and $\mathcal{G}$. Then*

$$Rad^{abs}(\mathcal{F}_{\alpha,\theta} + \mathcal{G}_{\beta,\theta}) \leq Rad^{abs}(\mathcal{F}_{\alpha,\theta}) + Rad^{abs}(\mathcal{G}_{\beta,\theta})$$

Lemma 5. *Let $\mathcal{F}$ be function class. L is some fixed real number and α is a real variable satisfying $|\alpha| \leq L$. Then*

$$Rad^{abs}(\alpha\mathcal{F}) \leq LRad^{abs}(\mathcal{F})$$

Lemma 6. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions and $\phi(0)$. Then, for any function class $\mathcal{F}$ of real-valued functions, the following inequality holds:*

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(f(x_i))| \right] \leq 2L \cdot \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)| \right] \tag{3}$$

Specifically, if g is a 1-Lipschitz activation function (such as ReLU) satisfying $g(0) = 0$, we have:

$$Rad^{abs}(g \circ \mathcal{F}) \leq 2Rad^{abs}(\mathcal{F}) \tag{4}$$

Proposition 1. *Let $\mathcal{F}$ be a real function class. Then we have:*

$$Rad^{abs}(relu \circ \mathcal{F}) \leq \frac{3}{2} Rad^{abs}(\mathcal{F})$$

4 Analysis on feed forward neural networks

A d-layer MLP mapping $\mathbb{R}^n$ to $\mathbb{R}$ can be written as:$NN(x) = f_d(g(f_{d-1}(\ldots\ldots g(f_0(\mathbf{x})))))$, where g is activation function satisfying 1-Lipschitz, $g(0) = 0$,$f_k(\mathbf{x}) = W_k\mathbf{x} + b_k$,

We define function class of neural network as following:

$$Neural = \{NN || W_k|_{i,j} \leq \frac{(A_k)_{i,j}}{2}, |b_k|_i \leq |B_k|_i\}$$

such that we restrict norm on each element. Then we claim following main result on feed-forward neural networks:

Theorem 2. *For sampling $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ where input dimension is n, we denote $R_i = \sup_k |\boldsymbol{x}_k|_i$*

$$Rad(Neural) \leq f_d(\cdots f_0(\begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \frac{R_2}{\sqrt{m}} \\ \vdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix}) \cdots) \text{ where } f_k(\boldsymbol{x}) = A_k\boldsymbol{x} + \frac{B_k}{\sqrt{m}}$$

The key idea behind Theorem 2 is to propagate the coordinate-wise Rademacher complexity through the network. The Lipschitz property of the activation allows each nonlinear layer to be controlled by the corresponding linear transformation, while the coordinate-wise constraints on the weights and biases lead to a deterministic comparison system. Iterating this comparison across layers yields the recursion

$$z_{k+1} = A_k z_k + \frac{B_k}{\sqrt{m}},$$

which directly gives the stated bound at the output layer. The complete proof is deferred to appendix.

Remark 1. *Theorem 2 is stated in a general form by assuming only that the activation function g is 1-Lipschitz and satisfies $g(0) = 0$. If we specialize to the ReLU activation, the contraction constant can be improved to $3/2$. Consequently, the assumption on the weight matrices can be relaxed from $|W_k|_{i,j} \leq \frac{(A_k)_{i,j}}{2}$ to $|W_k|_{i,j} \leq \frac{2(A_k)_{i,j}}{3}$.*

In the remainder of the paper, unless otherwise specified, we assume the ReLU activation and use the improved constant $3/2$. For a general 1-Lipschitz activation satisfying $g(0) = 0$, the corresponding constant is 2.

4.1 Residual neural networks

Residual networks (ResNets), introduced in 2015(He et al. (2016)), have become one of the most influential neural network architectures. The main improvement is the introduction of a residual connection in each layer: from $relu(Wx+b)$ to $x+relu(Wx+b)$. A complete treatment of practical ResNet architectures, including pooling layers in CNNs and attention mechanisms in Transformers, would considerably complicate the analysis. For convenience, we make following assumptions on ResNet architecture:

$$Res(\mathbf{x}) = f_d(f_{d-1} \cdots f_0(x))$$

Where $f_i(\mathbf{x}) = \mathbf{x} + g(W_i\mathbf{x} + b_i)$ for $i < d$. The final layer is $f_d(\mathbf{x}) = \mathbf{x}_k + g(W_d\mathbf{x} + b)$. Here $\mathbf{x}_k$ represents kth component of $\mathbf{x}$.For simplicity, we assume the residual branch has the same dimension as the skip connection and omit the projection matrix. The extension to a general linear projection is straightforward.

We claim following inequality:

Theorem 3. *For sampling $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ where input dimension is n, under assumption $|W_k|_{i,j} \leq \frac{2(A_k)_{i,j}}{3}, |b_k|_i \leq \frac{2(B_k)_i}{3}$, activation function $g(x) = Relu(x)$, $Rad^{abs}(Resnet) \leq f_d(\cdots f_0(Rad^{abs}(X))\cdots)$, where*

$Rad^{abs}(X) = \begin{pmatrix} Rad^{abs}(x_1) \\ \cdots \\ Rad^{abs}(x_n) \end{pmatrix}$, $f_0(\boldsymbol{x}) = A_0\boldsymbol{x} + \frac{B_0}{\sqrt{m}}$, $f_k(\boldsymbol{x}) = (I + A_k)\boldsymbol{x} + \frac{B_k}{\sqrt{m}}$ *for* $0 < k < d$ *and* $f_d(\boldsymbol{x}) = (e_k^T + A_k)x + \frac{B_k}{\sqrt{m}}$.

The resulting comparison system has the same coordinate-wise structure as that of feed-forward networks, with one important modification induced by the residual connection. In particular, each residual block transforms the comparison vector according to $z_{k+1} \leq (I + A_k)z_k + \frac{B_k}{\sqrt{m}}$,, where the identity matrix I corresponds directly to the skip connection. Thus, the effect of the residual architecture is explicitly preserved in the complexity bound rather than being absorbed into a generic norm estimate. This provides a unified way to analyze residual and feed-forward architectures through the same coordinate-wise comparison framework.

5 Analysis on Neural ODE

Neural ordinary differential equations (Neural ODEs) provide a continuous-depth formulation of deep neural networks. The key idea is to interpret the hidden state transformation of a neural network as a continuous dynamical system. Instead of stacking a finite number of residual blocks,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + g(W_k\mathbf{x}_k + b_k),$$

Neural ODEs consider the continuous limit of the residual connection: $\frac{d\mathbf{x}(t)}{dt} = g(W(t)\mathbf{x}(t) + b(t))$.

This formulation allows the depth of the network to be replaced by a continuous time variable, providing a flexible framework for modeling complex transformations. Moreover, many architectures can be interpreted from the Neural ODE perspective. For example, a ResNet block can be viewed as an Euler discretization of the corresponding differential equation. In this work, we focus on dynamical system:

$$\frac{d\mathbf{x}(t)}{dt} = g(W(t)\mathbf{x}(t) + b(t)),$$

which allows us to derive an explicit complexity propagation bound. We define the model as

$$F(\mathbf{x}(0); W(t), b(t), w) = w^T\mathbf{x}(1),$$

where $\mathbf{x}(1)$ is the solution of the ODE at time $t = 1$. Assume that w is a constant vector and $W(t)$ and $b(t)$ are model parameters that are continuous on $[0, 1]$ and satisfy $|W(t)|_{i,j} \leq A_{i,j}(t), |b(t)|_i \leq B_i(t)$, where $A(t), B(t)$ also change continuous over t. The corresponding Rademacher complexity can be analyzed by studying the propagation of the complexity of the hidden state along the continuous flow.

First we claim following result:

Theorem 4. *Let input data be $x_1, ...x_n$. Let $\mathcal{F}_1, \mathcal{F}_2, \cdots$ be a sequence of function class and $\mathcal{F}$ be a function class. If for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $||f - f'||_\infty \leq \varepsilon$ and for all $f \in \mathcal{F}$, for all $n > N$, there exists $f' \in \mathcal{F}_n$ such that $||f - f'||_\infty \leq \varepsilon$, we have*

$$\lim_{n\to\infty} Rad(\mathcal{F}_n) = Rad(\mathcal{F})$$

Proposition 2. *Let input data be $x_1, ...x_n$. Let $\mathcal{F}_1, \mathcal{F}_2, \cdots$ be a sequence of function class and $\mathcal{F}$ be a function class. If for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $||f - f'||_\infty \leq \varepsilon$ and for all $f \in \mathcal{F}$, for all $n > N$, there exists $f' \in \mathcal{F}_n$ such that $||f - f'||_\infty \leq \varepsilon$, we have*

$$\lim_{n\to\infty} Rad^{abs}(\mathcal{F}_n) = Rad^{abs}(\mathcal{F})$$

Proposition 3. *(Continuous-depth limit of ResNet). Consider the linear time-dependent ODE system $\frac{d\mathbf{x}(t)}{dt} = W(t)\mathbf{x}(t) + b(t), \quad \mathbf{x}(0) = \mathbf{x}_0$, where $W : [0,1] \to \mathbb{R}^{d\times d}$ and $b : [0,1] \to \mathbb{R}^d$ are continuous functions. Define the Euler discretization:*

$$\mathbf{x}_{k+1}^{(n)} = \left(I + \frac{W(k/n)}{n}\right)\mathbf{x}_k^{(n)} + \frac{b(k/n)}{n}, \quad k = 0, \ldots, n-1.$$

Equivalently, let $f_k^{(n)}(\mathbf{x}) = \left(I + \frac{W(k/n)}{n}\right)\mathbf{x} + \frac{b(k/n)}{n}$. Then

$$\lim_{n\to\infty} f_{n-1}^{(n)} \circ \cdots \circ f_0^{(n)}(\mathbf{x}_0) = \mathbf{x}(1).$$

Theorem 5 (Convergence of the Euler method, Hairer et al. (1993))**.** *Let $y(x)$ be the solution of the ODE $\frac{dy}{dx} = f(y,x), \quad y(x_0) = y_0$, where f satisfies a Lipschitz condition in y on a domain Γ with constant L. Let $\tilde{y}_n$ denote a numerical approximation obtained by an Euler-type method with initial error bounded by $\|y(x_0) - \tilde{y}_n(x_0)\| \leq K_n$, maximal step size bounded by H_n. Assume that as $n \to \infty, H_n \to 0, \quad K_n \to 0$. Then $\lim_{n\to\infty} \sup_x \|y(x) - \tilde{y}_n(x)\| = 0,$*

Final, our upper bound for Neural ODE yields as continuous path norm:

Theorem 6 (Aboslute Rademacher complexity bound for Neural ODE)**.** *For a sample $x_1, \ldots, x_m$, consider the Neural ODE system $\frac{d\mathbf{y}(t)}{dt} = g(W(t)\mathbf{y}(t)+b(t)), \quad \mathbf{y}(0) = x$.. Assume that $|W(t)|_{ij} \leq \frac{2A(t)_{ij}}{3}, \quad |b(t)|_i \leq \frac{2B(t)_i}{3}, \quad g(x) = Relu(x)$. Meanwhile $W(t)$ is C_W-Lipschitz continuous and $b(t)$ is C_B-Lipschitz continous. We assume $A(t), B(t)$ are continuous over t. Let $Y(x) = \mathbf{y}(1;x)$ denote the flow map of the ODE. Then for the hypothesis class $\mathcal{F} = \{x \mapsto w^T Y(x)\}$,, we have $Rad^{abs}(\mathcal{F}) \leq |w|^T Z(1)$,, where $Z(t)$ is the solution of the comparison system*

$$\frac{dZ(t)}{dt} = A(t)Z(t) + \frac{B(t)}{\sqrt{m}}, \qquad Z(0) = Rad^{abs}(X).$$

Thus, the proposed coordinate-wise comparison framework naturally extends from finite-depth residual networks to their continuous-depth limit. The resulting bound does not explicitly depend on the discretization depth, but is instead determined by the magnitude of the continuous parameter functions through the comparison system.

6 Analysis on attention block and transformer

We know that for given sequence input X, its attention is:

$$att(X) = softmax(XW_Q W_K^T X^T)XW_V^T$$

where $M = W_Q W_K^T, W_V$ are correspondingly query, key and value matrix.For simplicity of theoretical derivation and without loss of generality, we absorb the scaling factor $1/\sqrt{d_k}$ into the weight matrices W_Q and W_K. Moreover, we simplify as $M = W_Q W_K^T$. There is an intrinsic limit that dimension of $W_Q W_K^T$ does not exceed d_k, while we relax this limitation and complexity is non-decreasing. Then the expression of a transformer block can be expressed as following:

$B(X) = f_2(f_1(X))$, where $f_1(X) = X + att(X)(W_O X), f_2(X) = X + NN(X)$. Here $NN(X)$ is a two-layer neural network. Moreover, for a general transformer model, we can express it in this way: $transformer(X) = w^T B_n(B_{n-1} \cdots B_1(X)) + b$, where $B_1, \cdots, B_n$ are transformer blocks. In particular, when $n = 0$, transformer reduces to a linear model.

Consider single layer block $\mathcal{B} = \{f_2 \circ f_1 | f_1(X) = X + att(X)(W_O X), f_2(X) = W_2 g(W_1 X + b_1) + b_2, |M_{i,j}| \leq m_{i,j}, |(W_v)_{i,j}| \leq v_{i,j}, |(W_1)_{i,j}| \leq x_{i,j}, |(W_2)_{i,j}| \leq y_{i,j}\}$. We shall compute the upper bound for this single-layer transformer block class that we restrict on size of each parameter.

We claim following lemma:

Lemma 7 (Maurer's Vector Contraction Inequality)**.** *Let $\mathcal{X}$ be a state space and $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}^c$. Let $S = \{x_1, \ldots, x_n\}$ be a fixed sample of n points from $\mathcal{X}$.*

Assume that $h_i : \mathbb{R}^c \to \mathbb{R}$ for $i \in \{1, \ldots, n\}$ is a sequence of functions, and each h_i is L-Lipschitz continuous with respect to the ℓ_2-norm and $h_i(\mathbf{0}) = 0$. Then, for any independent Rademacher sequences $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\varepsilon' = (\varepsilon_{1,1}, \ldots, \varepsilon_{n,c})$, the following inequality holds:

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) \right] \leq \sqrt{2} L \cdot \mathbb{E}_{\varepsilon'} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right]$$

,

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i))| \right] \leq 2\sqrt{2} L \cdot \mathbb{E}_{\varepsilon'} \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i)| \right]$$

where $f_j(x_i)$ denotes the j-th component of the vector output $f(x_i)$.

Proposition 4. *Let $\mathcal{F}, \mathcal{G}$ be two function classes which are uniformly bounded and $\mathcal{H} = \{fg | f \in \mathcal{F}, g \in \mathcal{G}\}$. Then $Rad(\mathcal{H}) \leq 2(||\mathcal{G}||_\infty Rad(\mathcal{F}) + ||\mathcal{F}||_\infty Rad(\mathcal{G}))$*

Lemma 8. *Let X be a matrix of function classes where for all i, j, $0 \in X_{i,j}$ and $X_{i,j} = -X_{i,j}$, and A be a matrix of constant function class where $|A_{i,j}| \leq a_{i,j}$. Let $Rad^{abs}(X) = (Rad^{abs}(X_{i,j})), A' = (a_{i,j})$. Assuming XA is well defined, then we have:*

$$Rad^{abs}((XA)_{i,j}) \leq (Rad^{abs}(X)A')_{i,j}$$

And assuming AX is well defined, then we have:

$$Rad^{abs}((AX)_{i,j}) \leq (A' Rad^{abs}(X))_{i,j}$$

Proposition 5. *Let $\mathcal{F}, \mathcal{G}$ are tow function classse which are uniformly bounded and symmetric where $\mathcal{F} = -\mathcal{F}$ or $\mathcal{G} = -\mathcal{G}$. Let $\mathcal{H} = \{fg | f \in \mathcal{F}, g \in \mathcal{G}\}$. Then $Rad^{abs}(\mathcal{H}) \leq 2(||\mathcal{G}|| Rad^{abs}(\mathcal{F}) + ||\mathcal{F}|| Rad^{abs}(\mathcal{G}))$*

Proposition 6. *Let $softmax_i : (x_1, \cdots, x_n) \mapsto \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_i}}$. It is $\frac{\sqrt{2}}{4}$-Lipschitz continuous.*

Lemma 9. *Let $\mathcal{A}, \mathcal{B}$ be two matrices of function classes and A, B be two real matrices that for all $i, j, ||\mathcal{A}_{i,j}||_\infty \leq \frac{A_{i,j}}{2}, ||\mathcal{B}_{i,j}||_\infty \leq \frac{B_{i,j}}{2}$. Then we have: $Rad^{abs}(AB) \leq A Rad^{abs}(\mathcal{B}) + B Rad^{abs}(\mathcal{A})$. This inequality is understood as element-wise inequality.*

Theorem 7. *Let $h(X, Y, Z) = softmax(\frac{XY^T}{\sqrt{d_k}})ZW_O$ where X, Y, Z are matrices and above multiplication is well defined(softmax refers to row-wise softmax function) and d_k refers to size of row vector of X. For a transformer block function class*

$$B(X) = \{w^T f_1(f_2(X)) + b | f_1(X) = X + W_2 g(W_1 X + B_1) + B_2, f_2(X) = X + h(XW_K, XW_Q, XW_V)\}$$

under following norm restrictions:

$|W_2|_{i,j} \leq (Y_2)_{i,j}, |W_1|_{i,j} \leq (Y_1)_{i,j}, |b_2|_{i,j} \leq (B_2)_{i,j}, |b_1|_{i,j} \leq (B_1)_{i,j}, (W_Q)_{i,j} \leq Q_{i,j}, (W_K)_{i,j} \leq K_{i,j}, (W_V)_{i,j} \leq V_{i,j}, (XW_Q)_{i,j} \leq \frac{Q'_{i,j}}{2}, (XW_K)_{i,j} \leq \frac{K'_{i,j}}{2}, (XW_V)_{i,j} \leq \frac{V'_{i,j}}{2}, |W_O|_{i,j} \leq O_{i,j}$. *Then we have: $Rad^{abs}|(B) \leq |w|^T f_2(f_1(Rad^{abs}(X)))$, where*

$f_2(X) = X + Y_2(Y_1X + \frac{B_1}{\sqrt{m}}) + \frac{B_2}{\sqrt{m}}$, $f_1(X) = X + [\mathbf{1}_{n\times n}XV + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \frac{1}{2\sqrt{d_k}}\mathbf{1}_{n\times n}(XQK'^T + Q'(XK)^T)V']O$

We can apply this result iteratively to derive an upper bound on the generalization error of the full Transformer. The generalized result is as following:

Theorem 8. *Let $h(X, Y, Z) = \text{softmax}\left(\frac{XY^T}{\sqrt{d_k}}\right)ZW_O$ where X, Y, Z are matrices, the multiplication is well-defined (softmax refers to row-wise softmax function), and d_k refers to the column dimension of X and Y. Consider an L-layer Transformer network function class:*

$$\mathcal{T}_L(X) = \left\{w^T\mathcal{B}_L \circ \mathcal{B}_{L-1} \circ \cdots \circ \mathcal{B}_1(X) + b\right\}$$

where each transformer block $\mathcal{B}_l(X)$ is defined as the composition of an MLP and an Attention layer:

$$\mathcal{B}_l(Z) = f_{l,MLP}(f_{l,attn}(Z))$$

with $f_{l,attn}(Z) = Z + h(ZW_{l,K}, ZW_{l,Q}, ZW_{l,V})$ and $f_{l,MLP}(Z) = Z + W_{l,2}g(W_{l,1}Z + B_{l,1}) + B_{l,2}$.

Assume the hypothesis space is restricted by the following matrix norm bounds for each layer $l \in [1, L]$: $|W_{l,2}| \le \frac{2Y_{l,2}}{3}, |W_{l,1}| \le \frac{2Y_{l,1}}{3}, |b_{l,2}| \le \frac{2B_{l,2}}{3}, |b_{l,1}| \le \frac{2B_{l,1}}{3}$, $|W_{l,Q}| \le Q_l, |W_{l,K}| \le K_l, |W_{l,V}| \le V_l$, $|Z_{l-1}W_{l,Q}| \le \frac{Q_l'}{2}, |Z_{l-1}W_{l,K}| \le \frac{K_l'}{2}, |Z_{l-1}W_{l,V}| \le \frac{V_l'}{2}$, $|W_{l,O}| \le O_l$.

Then the empirical Rademacher complexity of the network is bounded by:

$$Rad(\mathcal{T}_L) \le |w|^T\mathcal{M}_L \circ \mathcal{A}_L \circ \mathcal{M}_{L-1} \circ \mathcal{A}_{L-1} \cdots \circ \mathcal{M}_1 \circ \mathcal{A}_1\left(Rad^{abs}(X)\right)$$

where the complexity mapping operators for the MLP layer ($\mathcal{M}_l$) and the Attention layer ($\mathcal{A}_l$) are defined element-wise as:

For the MLP Layers (all $l \in [1, L]$): $\mathcal{M}_l(Z) = Z + Y_{l,2}\left(Y_{l,1}Z + \frac{B_{l,1}}{\sqrt{m}}\right) + \frac{B_{l,2}}{\sqrt{m}}$

For the Attention Layers: $\mathcal{A}_l(Z) = Z + \left[\mathbf{1}_{n\times n}ZV_l + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}\left(ZQ_lK_l'^T + Q_l'(ZK_l)^T\right)V_l'\right]O_l$

6.1 Empirical Validation

We empirically evaluate whether the proposed operator-based complexity bounds capture variations in the generalization performance of neural networks. We conduct teacher-student regression experiments with randomly generated student models, considering MLPs, ResNets, and Transformer. For each architecture, we generate a collection of models with varying hyperparameters and examine the rank correlation between the proposed bound and the test MSE.

The results show a consistent correlation between the proposed bounds and generalization performance across the considered architectures, suggesting that the coordinate-wise complexity captures meaningful variations in generalization behavior. Additional experiments further examine the relationship between the bound, sample size, and the generalization gap.

Detailed experimental settings, including data generation, model architectures, hyperparameter ranges, training procedures, and additional results, are provided in Appendix B .

7 Conclusion

In this work, we introduced a unified framework for deriving generalization bounds based on coordinatewise comparison systems and absolute Rademacher complexity. The framework naturally incorporates biases and does not rely on positive homogeneity, while providing a common approach to analyze MLPs, ResNets, Transformers, and Neural ODEs. Our theoretical results establish path-norm-based bounds for both discrete and continuous architectures, and our experiments demonstrate a consistent correlation between the proposed complexity bounds and empirical generalization performance. These results suggest that coordinatewise comparison provides a useful and broadly applicable perspective for understanding the generalization behavior of modern neural networks.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and improving the clarity and presentation of the Abstract, Introduction, and Conclusion sections, as well as for discussing the feasibility and potential directions of research ideas.

We have not used generative AI tools to generate mathematical results, mathematical statements, proofs, or derivations, and these uses are not applicable to this work.

Additionally, we used generative AI tools for brainstorming and discussing research ideas. We have reviewed and independently verified all AI-assisted work. In particular, all mathematical results, proofs, derivations, and technical claims were developed and verified by the authors independently of generative AI. The authors take full responsibility for the final content of this work, including all text, claims, mathematical results, and other artifacts produced with the aid of generative AI.

REFERENCES

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

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms, 2022. URL https://arxiv.org/abs/2110.10090.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. doi: 10.1162/neco.1992.4.1.1.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 2018.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2 edition, 1993.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *The Twelfth International Conference on Learning Representations*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

Haotian Ju, Dongyue Li, and Hongyang R. Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10481–10500. PMLR, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Fabian Latorre, Antoine Bonnet, Paul Rolland, Nadav Hallak, and Volkan Cevher. 1-path-norm regularization of deep neural networks. In *LatinX in AI Workshop at the 40th International Conference on Machine Learning*, 2023.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Berlin, Heidelberg, 1991.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1352–1368, 2021.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 2023.

Fanghui Liu, Leello Dadi, and Volkan Cevher. Learning with norm constrained, over-parameterized, two-layer neural networks. *Journal of Machine Learning Research*, 2024.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818. PMLR, 2024.

Pierre Marion. Generalization bounds for neural ordinary differential equations and deep residual networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48918–48938, 2023.

Andreas Maurer. A vector-contraction inequality for rademacher complexities, 2016. URL `https://arxiv.org/abs/1605.00251`.

Brian Mwigo and Anirban Dasgupta. Generalization bound for a shallow transformer trained using gradient descent. *Transactions on Machine Learning Research*, 2026.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Annual Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401. PMLR, 2015.

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

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR, 2024.

Madhusudan Verma and Manoj Kumar. Generalization bound for a general class of neural ordinary differential equations, 2025. URL `https://arxiv.org/abs/2508.18920`.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2020.

Xinhao Yao, Hongjin Qian, Xiaolin Hu, Gengze Xu, Wei Liu, Jian Luan, Bin Wang, and Yong Liu. Theoretical insights into fine-tuning attention mechanism: Generalization and optimization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6830–6838, 2025.

A PROOF

We provide the proofs of the theorems, propositions, and lemmas in the main text.

Lemma 1. Let $\mathcal{F}_{\alpha,\theta}$ be parametrized function class depending on α and θ; $\mathcal{G}_{\beta,\theta}$ be parametrized function class depending on β, θ. α, β, θ are restricted in compact domains. θ are common parameters shared by $\mathcal{F}$ and $\mathcal{G}$. Then

$$Rad(\mathcal{F}_{\alpha,\theta} + \mathcal{G}_{\beta,\theta}) \leq Rad(\mathcal{F}_{\alpha,\theta}) + Rad(\mathcal{G}_{\beta,\theta})$$

Proof. By the definition of empirical Rademacher complexity for a fixed sample $S = \{x_1, \ldots, x_n\}$, we have:

$$\begin{aligned} Rad(\mathcal{F} + \mathcal{G}) &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f_{\alpha,\theta}(x_i) + g_{\beta,\theta}(x_i)) \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i) \right) \right] \end{aligned}$$

Using the property that the supremum of a sum is less than or equal to the sum of the suprema, i.e., $\sup(A + B) \leq \sup A + \sup B$, we get:

$$\begin{aligned} Rad(\mathcal{F} + \mathcal{G}) &\leq \mathbb{E}_\varepsilon \left[\sup_{\alpha,\theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i) + \sup_{\beta,\theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i) \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i) \right] + \mathbb{E}_\varepsilon \left[\sup_{\beta,\theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i) \right] \\ &= Rad(\mathcal{F}_{\alpha,\theta}) + Rad(\mathcal{G}_{\beta,\theta}) \end{aligned}$$

This completes the proof. □

Lemma 2. Let $\mathcal{F}$ be function class where $0 \in \mathcal{F}$. L is some fixed real number and α is a real variable satisfying $|\alpha| \leq \frac{L}{2}$. Then

$$Rad(\alpha\mathcal{F}) \leq L Rad(\mathcal{F})$$

Proof. By the definition of empirical Rademacher complexity for a fixed sample $S = \{x_1, \ldots, x_n\}$:

$$Rad(\alpha\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}, |\alpha| \leq \frac{L}{2}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \alpha f(x_i) \right]$$

For any fixed realization of Rademacher variables ε and any function $f \in \mathcal{F}$, the expression inside the supremum is a linear function of α. Let $X_f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)$. The maximum of the linear function $\phi(\alpha) = \alpha X_f$ over the compact interval $[-\frac{L}{2}, \frac{L}{2}]$ must be attained at one of the endpoints:

$$\sup_{|\alpha| \leq \frac{L}{2}} \alpha X_f = \max \left(\frac{L}{2} X_f, -\frac{L}{2} X_f \right) = \frac{L}{2} |X_f|$$

Substituting this back into the expectation, we have:

$$Rad(\alpha\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{L}{2} |X_f| \right] = \frac{L}{2} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \max(X_f, -X_f) \right]$$

Observing that $\sup_{f \in \mathcal{F}} \max(X_f, -X_f)$ is equivalent to taking the supremum over the union of the original function class $\mathcal{F}$ and its negation $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$, we obtain:

$$Rad(\alpha\mathcal{F}) = \frac{L}{2} Rad(\mathcal{F} \cup -\mathcal{F}) \tag{5}$$

While we know

$$\begin{aligned} Rad(\mathcal{F} \cup -\mathcal{F}) &= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F} \cup -\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right] \\ &= \mathbb{E}_\varepsilon \left[\max\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i), \sup_{f \in -\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \} \right] \\ &\leq \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) + \sup_{f \in -\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right] \\ &= Rad(\mathcal{F}) + Rad(-\mathcal{F}) = 2Rad(\mathcal{F}) \end{aligned}$$

The inequality comes from $\max\{a, b\} \leq a + b$ when a, b are nonnegative. While $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i), \sup_{f \in -\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)$ are always nonnegative by $0 \in \mathcal{F}$.

Combining previous results, we have:

$$Rad(\alpha \mathcal{F}) \leq L Rad(\mathcal{F})$$

□

Lemma 4. Let $\mathcal{F}_{\alpha,\theta}$ be parametrized function class depending on α and θ; $\mathcal{G}_{\beta,\theta}$ be parametrized function class depending on β, θ. α, β, θ are restricted in compact domains. θ are common parameters shared by $\mathcal{F}$ and $\mathcal{G}$. Then

$$Rad^{abs}(\mathcal{F}_{\alpha,\theta} + \mathcal{G}_{\beta,\theta}) \leq Rad^{abs}(\mathcal{F}_{\alpha,\theta}) + Rad^{abs}(\mathcal{G}_{\beta,\theta})$$

Proof. By the definition of empirical Rademacher complexity for a fixed sample $S = \{x_1, \ldots, x_n\}$, we have:

$$\begin{aligned} Rad^{abs}(\mathcal{F} + \mathcal{G}) &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f_{\alpha,\theta}(x_i) + g_{\beta,\theta}(x_i))| \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i) \right| \right] \end{aligned}$$

Using the property that the supremum of a sum is less than or equal to the sum of the suprema, i.e., $\sup |A + B| \leq \sup |A| + \sup |B|$, we get:

$$\begin{aligned} Rad^{abs}(\mathcal{F} + \mathcal{G}) &\leq \mathbb{E}_\varepsilon \left[\sup_{\alpha,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i)| + \sup_{\beta,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i)| \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i)| \right] + \mathbb{E}_\varepsilon \left[\sup_{\beta,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i)| \right] \\ &= Rad^{abs}(\mathcal{F}_{\alpha,\theta}) + Rad^{abs}(\mathcal{G}_{\beta,\theta}) \end{aligned}$$

This completes the proof. □

Lemma 5. Let $\mathcal{F}_{\alpha,\theta}$ be parametrized function class depending on α and θ; $\mathcal{G}_{\beta,\theta}$ be parametrized function class depending on β, θ. α, β, θ are restricted in compact domains. θ are common parameters shared by $\mathcal{F}$ and $\mathcal{G}$. Then

$$Rad^{abs}(\mathcal{F}_{\alpha,\theta} + \mathcal{G}_{\beta,\theta}) \leq Rad^{abs}(\mathcal{F}_{\alpha,\theta}) + Rad^{abs}(\mathcal{G}_{\beta,\theta})$$

Proof. By the definition of empirical Rademacher complexity for a fixed sample $S = \{x_1, \ldots, x_n\}$, we have:

$$\begin{aligned} Rad^{abs}(\mathcal{F} + \mathcal{G}) &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f_{\alpha,\theta}(x_i) + g_{\beta,\theta}(x_i))| \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\alpha,\beta,\theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\alpha,\theta}(x_i) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g_{\beta,\theta}(x_i) \right| \right] \end{aligned}$$

Using the property that the supremum of a sum is less than or equal to the sum of the suprema, i.e., $\sup|A+B| \leq \sup|A| + \sup|B|$, we get:

$$\begin{aligned} Rad^{abs}(\mathcal{F}+\mathcal{G}) &\leq \mathbb{E}_\varepsilon\left[\sup_{\alpha,\theta}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_{\alpha,\theta}(x_i)| + \sup_{\beta,\theta}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g_{\beta,\theta}(x_i)|\right] \\ &= \mathbb{E}_\varepsilon\left[\sup_{\alpha,\theta}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_{\alpha,\theta}(x_i)|\right] + \mathbb{E}_\varepsilon\left[\sup_{\beta,\theta}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g_{\beta,\theta}(x_i)|\right] \\ &= Rad^{abs}(\mathcal{F}_{\alpha,\theta}) + Rad^{abs}(\mathcal{G}_{\beta,\theta}) \end{aligned}$$

This completes the proof. □

Lemma 6. Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions and $\phi(0)$. Then, for any function class $\mathcal{F}$ of real-valued functions, the following inequality holds:

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i))|\right] \leq 2L\cdot\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)|\right] \tag{6}$$

Specifically, if g is a 1-Lipschitz activation function (such as ReLU) satisfying $g(0)=0$, we have:

$$Rad^{abs}(g\circ\mathcal{F}) \leq 2Rad^{abs}(\mathcal{F}) \tag{7}$$

Proof. Let $\mathcal{F}' = \mathcal{F}\cup\{0\}$, then we have:

$$\begin{aligned} Rad^{abs}(\phi\circ\mathcal{F}) &\leq Rad^{abs}(\phi\circ\mathcal{F}') \\ &= \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}'}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i))|\right] \\ &= \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}'}\max\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i)), -\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i))\}\right] \\ &\leq \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}'}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i)) + \sup_{f\in\mathcal{F}'}(-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(f(x_i)))\right] \\ &= 2Rad(\phi\circ\mathcal{F}') \\ &\leq 2LRad(\mathcal{F}') \\ &= 2L\mathbb{E}_\varepsilon\left[\max\{0, \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\}\right] \\ &\leq 2L\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)|\right] \\ &= 2LRad^{abs}(\mathcal{F}) \end{aligned}$$

□

Proposition 1. Let $\mathcal{F}$ be a real function class. Then we have:

$$Rad^{abs}(relu\circ\mathcal{F}) \leq \frac{3}{2}Rad^{abs}(\mathcal{F})$$

Proof. We have:

$$
\begin{aligned}
Rad^{abs}(relu \circ \mathcal{F}) &= Rad^{abs}(\frac{x+|x|}{2} \circ \mathcal{F}) \\
&\leq Rad^{abs}(\frac{x}{2} \circ \mathcal{F}) + Rad^{abs}(\frac{|x|}{2} \circ \mathcal{F}) \\
&\leq \frac{1}{2} Rad^{abs}(\mathcal{F}) + Rad^{abs}(\mathcal{F}) \\
&= \frac{3}{2} Rad^{abs}(\mathcal{F})
\end{aligned}
$$

□

Theorem 2. For sampling $\mathbf{x}_1, ..., \mathbf{x}_m$ where input dimension is n, we denote $R_i = \sup_k |\mathbf{x}_k|_i$

$$
Rad(Neural) \leq f_d(\cdots f_0(\begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \frac{R_2}{\sqrt{m}} \\ \vdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix}) \cdots)
$$

where

$$
f_k(\mathbf{x}) = A_k \mathbf{x} + \frac{B_k}{\sqrt{m}}
$$

Proof. We shall prove by induction. In fact, we strengthen our statement to following:

Let $r_k = \begin{pmatrix} r_{k,1} \\ \cdots \\ r_{k,d_k} \end{pmatrix}$ be the output of kth layer.we can prove that:

$\begin{pmatrix} Rad(r_{k,1}) \\ \cdots \\ Rad(r_{k,d_k}) \end{pmatrix} \leq A_k \begin{pmatrix} Rad(r_{k-1,1}) \\ \cdots \\ Rad(r_{k-1,d_{k-1}}) \end{pmatrix} + \frac{1}{\sqrt{m}} B_k$ for $k \geq 1$ and

$\begin{pmatrix} Rad(r_{0,1}) \\ \cdots \\ Rad(r_{0,d_1}) \end{pmatrix} \leq A_0 \begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \cdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix} + \frac{1}{\sqrt{m}} B_0$. Here inequality symbol represents element-wise inequality, such that for each corresponding element, inequality holds.

First we consider the first layer, we have:

$r_{1,k} = g(\sum_{i=1}^{n} (W_0)_{k,i} \mathbf{x}_i + b_{0,k})$.

Immediately, $Rad(r_{1,k}) \leq \sum_{i=1}^{n} Rad((W_0)_{k,i} \mathbf{x}_i) + Rad(b_0)$. By straightforward computation, we have $Rad((W_0)_{k,i} \mathbf{x}_i) \leq \frac{(A_0)_{k,i} R_i}{\sqrt{m}}$ and $Rad(b_0) \leq \frac{B_0}{\sqrt{m}}$.

Thus for each k, $Rad(r_{1,k}) \leq (A_0)_k \begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \frac{R_2}{\sqrt{m}} \\ \vdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix} + \frac{B_{0,k}}{\sqrt{m}}$.Here $(A_0)_k$ denotes the kth row of A_0.

By comparing each element, we have our result:

$\begin{pmatrix} Rad(r_{1,1}) \\ \cdots \\ Rad(r_{1,d_1}) \end{pmatrix} \leq A_0 \begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \cdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix} + \frac{1}{\sqrt{m}} B_0$.

Assuming our conclusion holds for layer k, then we consider $k+1$. We have

$r_{k+1,s} = g(\sum_{i=1}^{n}(W_{k+1})_{s,i}r_{k,i} + b_{k+1})$ By additivity and contraction lemma we have:

$$Rad(r_{k+1,s}) \leq \sum_{i=1}^{n} Rad((W_{k+1})_{s,i}r_{k,i}) + Rad(b_{k+1,s})$$
$$\leq \sum_{i=1}^{n}(A_{k+1})_{s,i}Rad(r_{k,i}) + \frac{B_{k+1,s}}{\sqrt{m}}$$
$$= (A_{k+1})_s \begin{pmatrix} Rad(r_{k,1}) \\ Rad(r_{k,2}) \\ \vdots \\ Rad(r_{k,n}) \end{pmatrix} + \frac{B_{k+1,s}}{\sqrt{m}}$$

By comparing elementwise, we have:

$$\begin{pmatrix} Rad(r_{k+1,1}) \\ \cdots \\ Rad(r_{k+1,d_{k+1}}) \end{pmatrix} \leq A_{k+1} \begin{pmatrix} Rad(r_{k,1}) \\ \cdots \\ Rad(r_{k,d_{k-1}}) \end{pmatrix} + \frac{1}{\sqrt{m}}B_{k+1}.$$

Thus the strengthened induction statement holds for all layers.

Now define

$$v_0 = \begin{pmatrix} \frac{R_1}{\sqrt{m}} \\ \frac{R_2}{\sqrt{m}} \\ \vdots \\ \frac{R_n}{\sqrt{m}} \end{pmatrix}.$$

From the first-layer estimate we have

$$\begin{pmatrix} Rad(r_{1,1}) \\ \cdots \\ Rad(r_{1,d_1}) \end{pmatrix} \leq f_0(v_0),$$

where

$$f_k(\mathbf{x}) = A_k\mathbf{x} + \frac{1}{\sqrt{m}}B_k.$$

Using the induction relation

$$\begin{pmatrix} Rad(r_{k,1}) \\ \cdots \\ Rad(r_{k,d_k}) \end{pmatrix} \leq f_k \left(\begin{pmatrix} Rad(r_{k-1,1}) \\ \cdots \\ Rad(r_{k-1,d_{k-1}}) \end{pmatrix} \right),$$

and repeatedly applying it, we obtain

$$\begin{pmatrix} Rad(r_{d,1}) \\ \cdots \\ Rad(r_{d,d_d}) \end{pmatrix} \leq f_d(f_{d-1}(\cdots f_1(f_0(v_0))\cdots)).$$

Since the network output is one-dimensional, we have

$$Rad(Neural) = Rad(r_{d,1}).$$

Taking the first component on both sides yields

$$Rad(Neural) \leq f_d(f_{d-1}(\cdots f_1(f_0(v_0))\cdots)),$$

which is exactly the desired conclusion.

□

Theorem 3. For sampling $\mathbf{x}_1, ..., \mathbf{x}_m$ where input dimension is n, under assumption $|W_k|_{i,j} \leq \frac{2(A_k)_{i,j}}{3}$, $|b_k|_i \leq \frac{2(B_k)_i}{3}$, activation function $g(x) = Relu(x)$

$Rad^{abs}(Resnet) \leq f_d(\cdots f_0(Rad^{abs}(X))\cdots)$, where

$Rad^{abs}(X) = \begin{pmatrix} Rad^{abs}(x_1) \\ \cdots \\ Rad^{abs}(x_n) \end{pmatrix}$, $f_0(\mathbf{x}) = A_0\mathbf{x} + \frac{B_0}{\sqrt{m}}$, $f_k(\mathbf{x}) = (I + A_k)\mathbf{x} + \frac{B_k}{\sqrt{m}}$ for $0 < k < d$ and $f_d(\mathbf{x}) = (e_k^T + A_k)x + \frac{B_k}{\sqrt{m}}$.

Proof. We prove the result by induction. In fact, we prove the stronger statement that for the output of the k-th residual block

$$r_k = \begin{pmatrix} r_{k,1} \\ \vdots \\ r_{k,d_k} \end{pmatrix},$$

the following inequality holds:

$$\begin{pmatrix} Rad^{abs}(r_{k,1}) \\ \vdots \\ Rad^{abs}(r_{k,d_k}) \end{pmatrix} \leq f_k \left(\begin{pmatrix} Rad^{abs}(r_{k-1,1}) \\ \vdots \\ Rad^{abs}(r_{k-1,d_{k-1}}) \end{pmatrix} \right),$$

where the inequality is understood element-wise.

For the first residual block, we have

$$r_{1,s} = x_s + g\left(\sum_{i=1}^{n} (W_0)_{s,i} x_i + b_{0,s}\right).$$

By the subadditivity of Rademacher complexity,

$$Rad^{abs}(r_{1,s}) \leq Rad^{abs}(x_s) + Rad^{abs}\left(g\left(\sum_i (W_0)_{s,i} x_i + b_{0,s}\right)\right).$$

By contraction lemma of relu function,

$$Rad^{abs}(r_{1,s}) \leq Rad^{abs}(x_s) + \frac{3}{2} Rad^{abs}\left(\sum_i (W_0)_{s,i} x_i + b_{0,s}\right).$$

Using additivity again,

$$Rad^{abs}(r_{1,s}) \leq Rad^{abs}(x_s) + \frac{3}{2}\left[\sum_i Rad^{abs}((W_0)_{s,i} x_i) + Rad^{abs}(b_{0,s}).\right]$$

By the scaling lemma and the definition of the constraint,

$$Rad^{abs}((W_0)_{s,i}x_i) \leq \frac{2}{3}(A_0)_{s,i}Rad^{abs}(x_i),$$

and

$$Rad^{abs}(b_{0,s}) \leq \frac{(B_0)_s}{\sqrt{m}}.$$

Therefore,

$$Rad^{abs}(r_{1,s}) \leq Rad^{abs}(x_s) + \sum_i (A_0)_{s,i}Rad^{abs}(x_i) + \frac{(B_0)_s}{\sqrt{m}}.$$

Hence the claim holds for the first layer.

Now assume the statement holds for the k-th layer. Consider the $(k+1)$-th residual block:

$$r_{k+1,s} = r_{k,s} + g\left(\sum_i (W_{k+1})_{s,i}r_{k,i} + b_{k+1,s}\right).$$

By subadditivity and the contraction lemma,

$$\begin{aligned} Rad^{abs}(r_{k+1,s}) &\leq Rad^{abs}(r_{k,s}) + Rad^{abs}\left(\frac{3}{2}\sum_i (W_{k+1})_{s,i}r_{k,i} + b_{k+1,s}\right) \\ &\leq Rad^{abs}(r_{k,s}) + \frac{3}{2}\sum_i Rad^{abs}((W_{k+1})_{s,i}r_{k,i}) + \frac{3}{2}Rad^{abs}(b_{k+1,s}). \end{aligned}$$

Applying the scaling lemma,

$$Rad^{abs}(r_{k+1,s}) \leq Rad^{abs}(r_{k,s}) + \sum_i (A_{k+1})_{s,i}Rad^{abs}(r_{k,i}) + \frac{(B_{k+1})_s}{\sqrt{m}}.$$

Thus,

$$Rad^{abs}(r_{k+1,s}) \leq ((I + A_{k+1})\begin{pmatrix} Rad^{abs}(r_{k,1}) \\ \vdots \\ Rad^{abs}(r_{k,d_k}) \end{pmatrix})_s + \frac{(B_{k+1})_s}{\sqrt{m}}.$$

Therefore,

$$\begin{pmatrix} Rad^{abs}(r_{k+1,1}) \\ \vdots \\ Rad^{abs}(r_{k+1,d_{k+1}}) \end{pmatrix} \leq (I + A_{k+1})\begin{pmatrix} Rad^{abs}(r_{k,1}) \\ \vdots \\ Rad^{abs}(r_{k,d_k}) \end{pmatrix} + \frac{B_{k+1}}{\sqrt{m}}.$$

Hence the induction hypothesis is preserved.

Finally, we consider the output layer separately. The final output is

$$f_d(r_d) = e_j^T r_d + g(W_d r_d + b_d).$$

By subadditivity and contraction lemma,

$$Rad^{abs}(f_d(r_d)) \leq Rad^{abs}(e_j^T r_d) + \frac{3}{2} Rad^{abs}(W_d r_d + b_d).$$

Therefore,

$$Rad^{abs}(f_d(r_d)) \leq Rad^{abs}(e_j^T r_d) + A_d \begin{pmatrix} Rad^{abs}(r_{d,1}) \\ \vdots \\ Rad^{abs}(r_{d,d_d}) \end{pmatrix} + \frac{B_d}{\sqrt{m}}.$$

Since

$$Rad^{abs}(e_j^T r_d) = e_j^T \begin{pmatrix} Rad^{abs}(r_{d,1}) \\ \vdots \\ Rad^{abs}(r_{d,d_d}) \end{pmatrix},$$

we obtain

$$Rad^{abs}(ResNet) \leq (e_j^T + A_d) \begin{pmatrix} Rad^{abs}(r_{d,1}) \\ \vdots \\ Rad^{abs}(r_{d,d_d}) \end{pmatrix} + \frac{B_d}{\sqrt{m}},$$

Hence the final layer corresponds to the transformation

$$f_d(x) = (e_j^T + A_d)x + \frac{B_d}{\sqrt{m}}.$$

Repeating the recursion gives

$$\begin{pmatrix} Rad^{abs}(r_{d,1}) \\ \vdots \\ Rad^{abs}(r_{d,d_d}) \end{pmatrix} \leq f_d \circ \cdots \circ f_0(Rad^{abs}(X))$$

Finally, since the output layer selects the j-th coordinate,

$$Rad^{abs}(ResNet) = Rad^{abs}(r_{d,j}),$$

and taking the corresponding component gives

$$Rad^{abs}(ResNet) \leq f_d \circ \cdots \circ f_0(Rad^{abs}(X))$$

This completes the proof. □

Theorem 4. Let input data be $x_1, ...x_n$. Let $\mathcal{F}_1, \mathcal{F}_2, \cdots$ be a sequence of function class and $\mathcal{F}$ be a function class. If for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $||f - f'||_\infty \leq \varepsilon$ and for all $f \in \mathcal{F}$, for all $n > N$, there exists $f' \in \mathcal{F}_n$ such that $||f - f'||_\infty \leq \varepsilon$, we have

$$\lim_{n \to \infty} Rad(\mathcal{F}_n) = Rad(\mathcal{F})$$

Proof. For a fixed sample $S = \{x_1, \ldots, x_m\}$ and fixed Rademacher variables $\sigma_1, \ldots, \sigma_m$, define the linear functional

$$L_\sigma(f) = \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i).$$

We first show that L_σ is Lipschitz continuous with respect to the uniform norm. For any two functions f, g,

$$\begin{aligned}|L_\sigma(f) - L_\sigma(g)| &= \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i) - g(x_i))\right| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}|f(x_i) - g(x_i)| \\ &\leq \|f - g\|_\infty.\end{aligned}$$

Therefore,

$$|L_\sigma(f) - L_\sigma(g)| \leq \|f - g\|_\infty.$$

Now fix $\varepsilon > 0$. By the assumption of the theorem, there exists N such that for all $n > N$,

$$\forall f_n \in \mathcal{F}_n, \exists f \in \mathcal{F}, \quad \|f_n - f\|_\infty \leq \varepsilon.$$

Hence for every $f_n \in \mathcal{F}_n$,

$$L_\sigma(f_n) \leq L_\sigma(f) + \varepsilon.$$

Taking supremum over $f_n \in \mathcal{F}_n$ gives

$$\sup_{f_n \in \mathcal{F}_n} L_\sigma(f_n) \leq \sup_{f \in \mathcal{F}} L_\sigma(f) + \varepsilon.$$

Similarly, using the other direction of the assumption,

$$\sup_{f \in \mathcal{F}} L_\sigma(f) \leq \sup_{f_n \in \mathcal{F}_n} L_\sigma(f_n) + \varepsilon.$$

Combining the two inequalities,

$$\left|\sup_{f_n \in \mathcal{F}_n} L_\sigma(f_n) - \sup_{f \in \mathcal{F}} L_\sigma(f)\right| \leq \varepsilon.$$

Taking expectation with respect to the Rademacher variables,

$$|Rad_S(\mathcal{F}_n) - Rad_S(\mathcal{F})| \leq \varepsilon.$$

Since ε is arbitrary, we conclude that

$$\lim_{n\to\infty} Rad_S(\mathcal{F}_n) = Rad_S(\mathcal{F}).$$

□

Proposition 2. Let input data be $x_1, ...x_n$. Let $\mathcal{F}_1, \mathcal{F}_2, \cdots$ be a sequence of function class and $\mathcal{F}$ be a function class. If for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $||f - f'||_\infty \leq \varepsilon$ and for all $f \in \mathcal{F}$, for all $n > N$, there exists $f' \in \mathcal{F}_n$ such that $||f - f'||_\infty \leq \varepsilon$, we have

$$\lim_{n\to\infty} Rad^{abs}(\mathcal{F}_n) = Rad^{abs}(\mathcal{F})$$

Proof. We have $Rad^{abs}(\mathcal{F}_n) = Rad(\mathcal{F}_n \cup -\mathcal{F}_n), Rad^{abs}(\mathcal{F}) = Rad(\mathcal{F} \cup -\mathcal{F})$. So we only need to show hypothesis for theorem 3 holding for $(\mathcal{F}_n \cup -\mathcal{F}_n), (\mathcal{F} \cup -\mathcal{F})$

Fix $\varepsilon > 0$. By the assumption of the proposition, there exists an N such that for all $n > N$:

1. For every $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $\|f - f'\|_\infty \leq \varepsilon$. 2. For every $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}_n$ such that $\|f - f'\|_\infty \leq \varepsilon$.

We now show that the same condition holds for the symmetrized function classes. Let $n > N$.

For the first direction, let $g' \in \mathcal{F}_n \cup -\mathcal{F}_n$.

- If $g' \in \mathcal{F}_n$, by condition 1, there exists $g \in \mathcal{F} \subset (\mathcal{F} \cup -\mathcal{F})$ such that $\|g - g'\|_\infty \leq \varepsilon$.
- If $g' \in -\mathcal{F}_n$, then $g' = -f'$ for some $f' \in \mathcal{F}_n$. By condition 1, there exists $f \in \mathcal{F}$ such that $\|f - f'\|_\infty \leq \varepsilon$. Let $g = -f$. Since $f \in \mathcal{F}$, we have $g \in -\mathcal{F} \subset (\mathcal{F} \cup -\mathcal{F})$. Furthermore, $\|g - g'\|_\infty = \|(-f) - (-f')\|_\infty = \| - (f - f')\|_\infty = \|f - f'\|_\infty \leq \varepsilon$.

In both cases, we find a $g \in \mathcal{F} \cup -\mathcal{F}$ such that $\|g - g'\|_\infty \leq \varepsilon$.

For the second direction, let $g \in \mathcal{F} \cup -\mathcal{F}$.

- If $g \in \mathcal{F}$, by condition 2, there exists $g' \in \mathcal{F}_n \subset (\mathcal{F}_n \cup -\mathcal{F}_n)$ such that $\|g - g'\|_\infty \leq \varepsilon$.
- If $g \in -\mathcal{F}$, then $g = -f$ for some $f \in \mathcal{F}$. By condition 2, there exists $f' \in \mathcal{F}_n$ such that $\|f - f'\|_\infty \leq \varepsilon$. Let $g' = -f'$. Since $f' \in \mathcal{F}_n$, we have $g' \in -\mathcal{F}_n \subset (\mathcal{F}_n \cup -\mathcal{F}_n)$. Furthermore, $\|g - g'\|_\infty = \|(-f) - (-f')\|_\infty = \|f - f'\|_\infty \leq \varepsilon$.

In both cases, we find a $g' \in \mathcal{F}_n \cup -\mathcal{F}_n$ such that $\|g - g'\|_\infty \leq \varepsilon$.

Since both directions of the uniform approximation hypothesis hold for the classes $(\mathcal{F}_n \cup -\mathcal{F}_n)$ and $(\mathcal{F} \cup -\mathcal{F})$, we can apply the result of Theorem 3:

$$\lim_{n\to\infty} Rad(\mathcal{F}_n \cup -\mathcal{F}_n) = Rad(\mathcal{F} \cup -\mathcal{F})$$

Which immediately yields:

$$\lim_{n\to\infty} Rad^{abs}(\mathcal{F}_n) = Rad^{abs}(\mathcal{F})$$

This completes the proof. □

Proposition 3.(Continuous-depth limit of ResNet). Consider the linear time-dependent ODE system

$$\frac{d\mathbf{x}(t)}{dt} = W(t)\mathbf{x}(t) + b(t), \quad \mathbf{x}(0) = \mathbf{x}_0,$$

where $W : [0, 1] \to \mathbb{R}^{d\times d}$ and $b : [0, 1] \to \mathbb{R}^d$ are continuous functions. Define the Euler discretization:

$$\mathbf{x}_{k+1}^{(n)} = \left(I + \frac{W(k/n)}{n}\right)\mathbf{x}_k^{(n)} + \frac{b(k/n)}{n}, \quad k = 0, \ldots, n-1.$$

Equivalently, let

$$f_k^{(n)}(\mathbf{x}) = \left(I + \frac{W(k/n)}{n}\right)\mathbf{x} + \frac{b(k/n)}{n}.$$

Then

$$\lim_{n\to\infty} f_{n-1}^{(n)} \circ \cdots \circ f_0^{(n)}(\mathbf{x}_0) = \mathbf{x}(1).$$

Proof. Define

$$F(t, \mathbf{x}) = W(t)\mathbf{x} + b(t).$$

The solution of the ODE satisfies the integral equation

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t F(s, \mathbf{x}(s))ds.$$

Let

$$t_k = \frac{k}{n},$$

and denote the numerical error by

$$\mathbf{e}_k = \mathbf{x}(t_k) - \mathbf{x}_k^{(n)}.$$

For the exact solution,

$$\mathbf{x}(t_{k+1}) = \mathbf{x}(t_k) + \int_{t_k}^{t_{k+1}} F(s, \mathbf{x}(s))ds.$$

For the Euler scheme,

$$\mathbf{x}_{k+1}^{(n)} = \mathbf{x}_k^{(n)} + \frac{1}{n}F(t_k, \mathbf{x}_k^{(n)}).$$

Therefore,

$$\begin{aligned}\mathbf{e}_{k+1} &= \mathbf{e}_k + \int_{t_k}^{t_{k+1}} F(s, \mathbf{x}(s))ds - \frac{1}{n}F(t_k, \mathbf{x}_k^{(n)}) \\ &= \mathbf{e}_k + \frac{1}{n}\left(F(t_k, \mathbf{x}(t_k)) - F(t_k, \mathbf{x}_k^{(n)})\right) + \mathbf{r}_k,\end{aligned}$$

where

$$\mathbf{r}_k = \int_{t_k}^{t_{k+1}} \left(F(s, \mathbf{x}(s)) - F(t_k, \mathbf{x}(t_k))\right) ds.$$

Since $W(t)$ is continuous on a compact set,

$$L = \max_{t\in[0,1]} \|W(t)\| < \infty.$$

Hence

$$\|F(t, \mathbf{x}) - F(t, \mathbf{y})\| = \|W(t)(\mathbf{x} - \mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|.$$

Thus,

$$\|\mathbf{e}_{k+1}\| \le \left(1 + \frac{L}{n}\right)\|\mathbf{e}_k\| + \|\mathbf{r}_k\|.$$

Because $F(t, \mathbf{x}(t))$ is continuous on $[0, 1]$, it is uniformly continuous. Therefore there exists a modulus of continuity $\omega(\delta)$ with

$$\omega(\delta) \to 0 \quad (\delta \to 0)$$

such that

$$\|\mathbf{r}_k\| \le \frac{1}{n}\omega(1/n).$$

Using the recursive inequality and $\mathbf{e}_0 = 0$, we obtain

$$\|\mathbf{e}_n\| \le \omega(1/n) \sum_{k=0}^{n-1} \frac{1}{n}\left(1 + \frac{L}{n}\right)^{n-1-k}.$$

The geometric sum gives

$$\|\mathbf{e}_n\| \le \omega(1/n)\frac{(1 + L/n)^n - 1}{L}.$$

Since

$$\left(1+\frac{L}{n}\right)^n \le e^L,$$

we have

$$\|\mathbf{e}_n\| \le C\omega(1/n) \to 0.$$

Therefore,

$$\mathbf{x}_n^{(n)} \to \mathbf{x}(1),$$

which proves

$$\lim_{n\to\infty} f_{n-1}^{(n)} \circ \cdots \circ f_0^{(n)}(\mathbf{x}_0) = \mathbf{x}(1).$$

□

Theorem 6.[Aboslute Rademacher complexity bound for Neural ODE]

For a sample $\{x_1, \ldots, x_m\}$, consider the Neural ODE system

$$\frac{d\mathbf{y}(t)}{dt} = g(W(t)\mathbf{y}(t) + b(t)), \quad \mathbf{y}(0) = x.$$

Assume that

$$|W(t)|_{ij} \le \frac{2A(t)_{ij}}{3}, \qquad |b(t)|_i \le \frac{2B(t)_i}{3} \qquad g(x) = Relu(x).$$

Meanwhile $W(t)$ is C_W-Lipschitz continuous and $b(t)$ is C_B-Lipschitz continous. We assume $A(t), B(t)$ are continuous over t. Let $Y(x) = \mathbf{y}(1; x)$ denote the flow map of the ODE. Then for the hypothesis class

$$\mathcal{F} = \{x \mapsto w^T Y(x)\},$$

we have

$$Rad^{abs}(\mathcal{F}) \le |w|^T Z(1),$$

where $Z(t)$ is the solution of the comparison system

$$\frac{dZ(t)}{dt} = A(t)Z(t) + \frac{B(t)}{\sqrt{m}},$$

with initial condition

$$Z(0) = Rad^{abs}(X)$$

Proof. We have $x(t)$ is C_X-Lipschitz for some fixed constant C_X. Only need to show $||\frac{dx}{dt}||_2$ is bounded, such that $g(W(t)x(t) + b(t))$ is bounded. By $W(t), b(t)$ is bounded, we only need to show $x(t)$ is bounded by some constant B_X. By

$||x(t)|| = ||x(0) + \int_0^t g(W(s)x(s) + b(s))ds||$
$\le ||x_0|| + \int_0^t (||W||||x(s)|| + ||b||)ds \le ||W|| \int_0^t ||x(s)||ds + ||x_0|| + ||b||.$

From Gronwall inequality, immediately we have $||x||$ is bounded. Thus x is C_X-Lipschitz continous for some constant C_X.

Consider function class series $\mathcal{F}_n$ be set of residual neural network with paramemter matrices: $(\frac{W(0)}{n}, \frac{B(0)}{n}); (\frac{W(\frac{1}{n})}{n}, \frac{B(\frac{1}{n})}{n}); \cdots; (\frac{W(1)}{n}, \frac{B(1)}{n})$.

First we show that such series $\mathcal{F}_i, \cdots$ "converges" to ODE, such that for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f' \in \mathcal{F}_n$, there exists $f \in \mathcal{F}$ such that $||w^T f - w^T f'||_\infty \leq \varepsilon$ and for all $f \in \mathcal{F}$, for all $n > N$, there exists $f' \in \mathcal{F}_n$ such that $||w^T f - w^T f'||_\infty \leq \varepsilon$. We prove the result by showing both directions of approximation.

First we notice that $||w^T f - w^T f'||_\infty \leq ||w||_2 || \sup_x ||f - f'||_2$. So our goal reduces to:$\sup_x ||f - f'||_2 \leq \frac{\varepsilon}{||w||_2}$. By taking $\varepsilon' = \frac{\varepsilon}{||w||_2}$, we need to show $\sup_x ||f - f'||_2 \leq \varepsilon'$.

We prove the result by showing both directions of approximation.

First, consider an arbitrary function $f \in \mathcal{F}_{ODE}$ generated by

$$\dot{x}(t) = g(W(t)x(t) + B(t)).$$

Let $x(t)$ be the solution trajectory.

The corresponding Euler discretization is

$$x_{k+1}^n = x_k^n + \frac{1}{n} g(W(t_k)x_k^n + B(t_k)),$$

where

$$t_k = \frac{k}{n}.$$

Define the error

$$e_k = |x_k^n - x(t_k)|.$$

We have

$$\begin{aligned} e_{k+1} &= \left| x_k^n + \frac{1}{n} g(W(t_k)x_k^n + B(t_k)) - x(t_{k+1}) \right| \\ &\leq |x_k^n - x(t_k)| \\ &+ \frac{1}{n} |g(W(t_k)x_k^n + B(t_k)) - g(W(t_k)x(t_k) + B(t_k))| \\ &+ \tau_k, \end{aligned}$$

where τ_k is the local truncation error:

$$\tau_k = \left| x(t_k) + \frac{1}{n} g(W(t_k)x(t_k) + B(t_k)) - x(t_{k+1}) \right|.$$

Since g is Lipschitz continuous and $W(t)$ is bounded on $[0, 1]$,

$$|g(W(t_k)x_k^n + B(t_k)) - g(W(t_k)x(t_k) + B(t_k))| \leq ||W|| e_k.$$

Moreover, by the Lipschitz continuity of $W(t)$, $B(t)$ and $x(t)$,obviously $g(W(t)x(t) + B(t))$ is C-Lipschitz continuous for some fixed constant C. We take out this constant.

$$\begin{aligned} \tau_k &= \left| \int_{t_k}^{t_{k+1}} \left(g(W(s)x(s) + b(s)) - g(W(t_k)x(t_k) + b(t_k)) \right) ds \right| \\ &\leq \int_{t_k}^{t_{k+1}} C|s - t_k| \, ds \\ &= \frac{C}{2n^2}. \end{aligned}$$

Therefore,

$$e_{k+1} \leq (1 + \frac{||W||}{n})e_k + \frac{C}{2n^2}.$$

Applying the discrete Gronwall inequality gives

$$e_n \leq \frac{C}{2n||W||}[(1 + \frac{||W||}{n})^n - 1] \leq \frac{C'}{n}.$$

Where

$$C' = \frac{[(1 + \frac{||W||}{n})^n - 1]}{2||W||}$$

Hence,

$$\sup_x |f_n(x) - f(x)|_2 \to 0.$$

Therefore, for every $\varepsilon > 0$, there exists N such that for all $n > N$, $f_n \in \mathcal{F}_n$

$$\inf_{f \in \mathcal{F}_{ODE}} \sup_x \|f_n(x) - f(x)\|_2 < \varepsilon.$$

Conversely, take any $f \in \mathcal{F}_{ODE}$. Choosing the discrete parameters

$$W_k = W(\frac{k}{n}), \qquad B_k = B(\frac{k}{n}),$$

constructs a residual network $f_n \in \mathcal{F}_n$.

By the Euler convergence result in theorem 5(easy to check all hypethesis holding),

$$\lim_{n\to\infty} \sup_x \|f_n - f\|_2 = 0.$$

Combining two sides, we show that $\mathcal{F}_\rangle$

By previous result on residual neural networks, we have:

$$Rad^{abs}(w^T \mathcal{F}_n) \leq |w|^T f_n \circ \cdots \circ f_0(Z(0))$$

where

$$f_k(x) = (I + \frac{A(\frac{k}{n})}{n})x + \frac{B(\frac{k}{n})}{n\sqrt{m}}$$

By taking $n \to \infty$ on both sides of the inequality, we have:

$$\lim_{n\to\infty} Rad^{abs}(w^t \mathcal{F}_n) \leq \lim_{n\to\infty} |w|^T f_n \circ \cdots \circ f_0(Z(0)) = |w|^T \lim_{n\to\infty} f_n \circ \cdots \circ f_0(Z(0))$$

While by theorem 3, left hand side is $Rad(\mathcal{F}_{ODE})$; while right hand side is $|w|^T Z(1)$ by theorem 4. Thus our inequality is proven.

□

Lemma 7.[Maurer's Vector Contraction Inequality] Let $\mathcal{X}$ be a state space and $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}^c$. Let $S = \{x_1, \ldots, x_n\}$ be a fixed sample of n points from $\mathcal{X}$.

Assume that $h_i : \mathbb{R}^c \to \mathbb{R}$ for $i \in \{1, \ldots, n\}$ is a sequence of functions, and each h_i is L-Lipschitz continuous with respect to the ℓ_2-norm and $h_i(\mathbf{0}) = 0$. Then, for any independent Rademacher sequences $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\varepsilon' = (\varepsilon_{1,1}, \ldots, \varepsilon_{n,c})$, the following inequality holds:

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) \right] \leq \sqrt{2} L \cdot \mathbb{E}_{\varepsilon'} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right]$$

,

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i))| \right] \leq 2\sqrt{2} L \cdot \mathbb{E}_{\varepsilon'} \left[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i)| \right]$$

where $f_j(x_i)$ denotes the j-th component of the vector output $f(x_i)$.

Proof. The first inequality is the main result of Maurer (2016). For second inequality, let $\mathbf{0}$ be the zero function mapping any input x to the zero vector $(0, \ldots, 0) \in \mathbb{R}^c$. Let $\mathcal{F}' = \mathcal{F} \cup \{\mathbf{0}\}$. Since $h_i(\mathbf{0}) = 0$ for all i, evaluating h_i on the zero function yields 0. Thus, the supremum over $\mathcal{F}'$ is guaranteed to be non-negative.

Define the absolute vector Rademacher complexity as $Rad^{abs}_{vec}(\mathcal{F}) = \mathbb{E}_{\varepsilon_{i,j}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right| \right]$. We then have:

$$\begin{aligned}
Rad^{abs}(h \circ \mathcal{F}) &\leq Rad^{abs}(h \circ \mathcal{F}') \\
&= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}'} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) \right| \right] \\
&= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}'} \max \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)), -\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) \right\} \right] \\
&\leq \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}'} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) + \sup_{f \in \mathcal{F}'} \left(-\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h_i(f(x_i)) \right) \right] \\
&= 2Rad(h \circ \mathcal{F}') \quad \text{(by symmetry of } \varepsilon_i \text{ and } -\varepsilon_i) \\
&\leq 2\sqrt{2} L \, \mathbb{E}_{\varepsilon_{i,j}} \left[\sup_{f \in \mathcal{F}'} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right] \quad \text{(by Maurer's Vector Contraction Lemma)} \\
&= 2\sqrt{2} L \, \mathbb{E}_{\varepsilon_{i,j}} \left[\max \left\{ 0, \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right\} \right] \\
&\leq 2\sqrt{2} L \, \mathbb{E}_{\varepsilon_{i,j}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \varepsilon_{i,j} f_j(x_i) \right| \right] \\
&= 2\sqrt{2} L \, Rad^{abs}_{vec}(\mathcal{F})
\end{aligned}$$

□

Proposition 4. Let $\mathcal{F}, \mathcal{G}$ be two function classes which are uniformly bounded and $\mathcal{H} = \{fg | f \in \mathcal{F}, g \in \mathcal{G}\}$. Then $Rad(\mathcal{H}) \leq 2(||\mathcal{G}||_\infty Rad(\mathcal{F}) + ||\mathcal{F}||_\infty Rad(\mathcal{G}))$

Proof. We have

$$fg = \frac{1}{4||\mathcal{F}||_\infty ||\mathcal{G}||_\infty} [(||\mathcal{F}||_\infty g + ||\mathcal{G}||_\infty f)^2 - (||\mathcal{F}||_\infty g - ||\mathcal{G}||_\infty f)^2]$$

.

Thus,

$$Rad(\mathcal{H}) \leq \frac{1}{4||\mathcal{F}||_\infty||\mathcal{G}||_\infty}(Rad((||\mathcal{F}||_\infty g + ||\mathcal{G}||_\infty f)^2) + (||\mathcal{F}||_\infty g - ||\mathcal{G}||_\infty f)^2)$$

.

By Talagrand contraction lemma, we have:

$$Rad((||\mathcal{F}||_\infty g + ||\mathcal{G}||_\infty f)^2) \leq 4||\mathcal{F}||_\infty||\mathcal{G}||_\infty Rad((||\mathcal{F}||_\infty g + ||\mathcal{G}||_\infty f))$$

.

Similarly,

$$Rad((||\mathcal{F}||_\infty g - ||\mathcal{G}||_\infty f)^2) \leq 4||\mathcal{F}||_\infty||\mathcal{G}||_\infty Rad((||\mathcal{F}||_\infty g + ||\mathcal{G}||_\infty f))$$

.

Above all,

$$\begin{aligned} \text{Rad}(\mathcal{H}) &\leq [\text{Rad}((\|\mathcal{F}\|_\infty g + \|\mathcal{G}\|_\infty f)) + \text{Rad}((\|\mathcal{F}\|_\infty g - \|\mathcal{G}\|_\infty f))] \\ &\leq 2\left(\|\mathcal{G}\|_\infty \text{Rad}(\mathcal{F}) + \|\mathcal{F}\|_\infty \text{Rad}(\mathcal{G})\right) \end{aligned} \tag{8}$$

□

Lemma 8. Let X be a matrix of function classes where for all i, j, $0 \in X_{i,j}$ and $X_{i,j} = -X_{i,j}$, and A be a matrix of constant function class where $|A_{i,j}| \leq a_{i,j}$. Let $Rad^{abs}(X) = (Rad^{abs}(X_{i,j}))$, $A' = (a_{i,j})$. Assuming XA is well defined, then we have:

$$Rad^{abs}((XA)_{i,j}) \leq (Rad^{abs}(X)A')_{i,j}$$

And assuming AX is well defined, then we have:

$$Rad^{abs}((AX)_{i,j}) \leq (A'Rad^{abs}(X))_{i,j}$$

Proof. for fixed i, j, we have

$$(XA)_{i,j} = \sum_k X_{i,j}A_{k,j}$$

.

Thus

$$Rad^{abs}((XA)_{i,j}) \leq \sum_k Rad^{abs}(X_{i,j}A_{k,j}) \leq \sum_k a_{k,j}Rad^{abs}(X_{i,k}) = (Rad^{abs}(X)A')_{i,j}$$

.

Above all our conclusion holds. Another inequality can be proven similarly.

□

Proposition 5. Let $\mathcal{F}, \mathcal{G}$ are tow function classse which are uniformly bounded and symmetric where $\mathcal{F} = -\mathcal{F}$ or $\mathcal{G} = -\mathcal{G}$. Let $\mathcal{H} = \{fg | f \in \mathcal{F}, g \in \mathcal{G}\}$. Then $Rad^{abs}(\mathcal{H}) \leq 2(||\mathcal{G}||Rad^{abs}(\mathcal{F}) + ||\mathcal{F}||Rad^{abs}(\mathcal{G}))$

Proof. Clearly $\mathcal{H} = -\mathcal{H}$, thus

$$\begin{aligned} Rad^{abs}(\mathcal{H}) &= Rad(\mathcal{H}) \\ &\leq 2(||\mathcal{G}||Rad(\mathcal{F}) + ||\mathcal{F}||Rad(\mathcal{G})) \\ &\leq 2(||\mathcal{G}||Rad^{abs}(\mathcal{F}) + ||\mathcal{F}||Rad^{abs}(\mathcal{G})) \end{aligned}$$

□

proposition 6. Let $softmax_i : (x_1, \cdots, x_n) \mapsto \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_i}}$. It is $\frac{\sqrt{2}}{4}$-Lipschitz continuous.

Proof. Clearly, the Lipschitz constant is bounded by $\sup_{\mathbf{x}} ||\nabla softmax_i||_2$. By direct computation, we have:

$$\nabla softmax_i = \begin{pmatrix} -softmax_i softmax_1 \\ \cdots \\ softmax_i(1 - softmax_i) \\ \cdots \\ -softmax_i softmax_n \end{pmatrix}$$

Thus

$$||\nabla softmax_i||_2^2 = softmax_i^2[(1 - softmax_i)^2 + \sum_{j \neq i} softmax_j^2]$$

.

Note that:

$$\sum_{j \neq i} softmax_j^2 \leq (\sum_{j \neq i} softmax_j)^2 = (1 - softmax_i)^2$$

.

Thus,

$$||\nabla softmax_i||_2^2 \leq 2softmax_i^2(1 - softmax_i)^2 \leq \frac{1}{8}$$

. Our desired inequality holds.

□

Lemma 9. Let $\mathcal{A}, \mathcal{B}$ be two matrices of function classes and A, B be two real matrices that for all $i, j, ||\mathcal{A}_{i,j}||_\infty \leq \frac{A_{i,j}}{2}, ||\mathcal{B}_{i,j}||_\infty \leq \frac{B_{i,j}}{2}$. Then we have:

$$Rad^{abs}(AB) \leq ARad^{abs}(\mathcal{B}) + BRad^{abs}(\mathcal{A})$$

.

This inequality is understood as element-wise inequality.

Proof. By the definition of matrix multiplication for function classes, the (i, j)-th element of the matrix $\mathcal{AB}$ is given by the summation (Minkowski sum) of products:

$$(\mathcal{AB})_{i,j} = \sum_k \mathcal{A}_{i,k}\mathcal{B}_{k,j}$$

Applying the Rademacher complexity operator to this (i, j)-th element, and utilizing the sub-additivity property of Rademacher complexity for hypothesis classes formed by summation ($Rad(\mathcal{F} + \mathcal{G}) \leq Rad(\mathcal{F}) + Rad(\mathcal{G})$), we obtain:

$$Rad^{abs}((\mathcal{AB})_{i,j}) = Rad^{abs}\left(\sum_k \mathcal{A}_{i,k}\mathcal{B}_{k,j}\right) \leq \sum_k Rad^{abs}(\mathcal{A}_{i,k}\mathcal{B}_{k,j})$$

For any two scalar function classes $\mathcal{F}$ and $\mathcal{G}$, the previous proposition states that:

$$Rad^{abs}(\mathcal{FG}) \leq 2\left(\|\mathcal{G}\|_\infty Rad^{abs}(\mathcal{F}) + \|\mathcal{F}\|_\infty Rad^{abs}(\mathcal{G})\right)$$

Applying this product rule to each individual term $\mathcal{A}_{i,k}\mathcal{B}_{k,j}$ in the summation, we have:

$$Rad^{abs}(\mathcal{A}_{i,k}\mathcal{B}_{k,j}) \leq 2\left(\|\mathcal{B}_{k,j}\|_\infty Rad^{abs}(\mathcal{A}_{i,k}) + \|\mathcal{A}_{i,k}\|_\infty Rad^{abs}(\mathcal{B}_{k,j})\right)$$

We are given the infinity norm bounds $\|\mathcal{A}_{i,k}\|_\infty \leq \frac{A_{i,k}}{2}$ and $\|\mathcal{B}_{k,j}\|_\infty \leq \frac{B_{k,j}}{2}$. Substituting these upper bounds into the inequality above yields:

$$\begin{aligned} Rad^{abs}(\mathcal{A}_{i,k}\mathcal{B}_{k,j}) &\leq 2\left(\frac{B_{k,j}}{2}Rad^{abs}(\mathcal{A}_{i,k}) + \frac{A_{i,k}}{2}Rad(\mathcal{B}_{k,j})\right) \\ &= Rad^{abs}(\mathcal{A}_{i,k})B_{k,j} + A_{i,k}Rad^{abs}(\mathcal{B}_{k,j}) \end{aligned}$$

Summing this upper bound over the index k, we get the total bound for the (i,j)-th element:

$$\begin{aligned} Rad^{abs}((\mathcal{AB})_{i,j}) &\leq \sum_k \left(Rad^{abs}(\mathcal{A}_{i,k})B_{k,j} + A_{i,k}Rad^{abs}(\mathcal{B}_{k,j})\right) \\ &= \sum_k A_{i,k}Rad^{abs}(\mathcal{B}_{k,j}) + \sum_k Rad^{abs}(\mathcal{A}_{i,k})B_{k,j} \end{aligned}$$

Recognizing that $\sum_k A_{i,k}Rad^{abs}(\mathcal{B}_{k,j})$ is exactly the definition of the (i,j)-th element of the matrix product $ARad^{abs}(\mathcal{B})$, and $\sum_k Rad^{abs}(\mathcal{A}_{i,k})B_{k,j}$ corresponds to the (i,j)-th element of the matrix product $Rad^{abs}(\mathcal{A})B$, we can write:

$$Rad^{abs}((\mathcal{AB})_{i,j}) \leq (ARad^{abs}(\mathcal{B}))_{i,j} + (Rad^{abs}(\mathcal{A})B)_{i,j}$$

Since this relationship holds for all indices i and j, the element-wise matrix inequality is established:

$$Rad^{abs}(\mathcal{AB}) \leq ARad^{abs}(\mathcal{B}) + Rad^{abs}(\mathcal{A})B$$

This completes the proof. □

Theorem 7. Let $h(X,Y,Z) = softmax(\frac{XY^T}{\sqrt{d_k}})ZW_O$ where X, Y, Z are matrices and above multiplication is well defined(softmax refers to row-wise softmax function) and d_k refers to size of row vector of X.

For a transformer block function class

$$B(X) = \{w^T f_1(f_2(X)) + b | f_1(X) = X + W_2 g(W_1 X + B_1) + B_2, f_2(X) = X + h(XW_K, XW_Q, XW_V)\}$$

under following norm restrictions:

$|W_2|_{i,j} \leq (Y_2)_{i,j}, |W_1|_{i,j} \leq (Y_1)_{i,j}, |b_2|_{i,j} \leq (B_2)_{i,j}, |b_1|_{i,j} \leq (B_1)_{i,j}, (W_Q)_{i,j} \leq Q_{i,j}, (W_K)_{i,j} \leq K_{i,j}, (W_V)_{i,j} \leq V_{i,j}, (XW_Q)_{i,j} \leq \frac{Q'_{i,j}}{2}, (XW_K)_{i,j} \leq \frac{K'_{i,j}}{2}, (XW_V)_{i,j} \leq \frac{V'_{i,j}}{2}, |W_O|_{i,j} \leq O_{i,j}$.

Then we have:

$$Rad^{abs}|(B) \leq |w|^T f_2(f_1(Rad^{abs}(X)))$$

, where

$f_2(X) = X + Y_2(Y_1 X + \frac{B_1}{\sqrt{m}}) + \frac{B_2}{\sqrt{m}}$, $f_1(X) = X + [\mathbf{1}_{n\times n}XV + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}(XQK'^T + Q'(XK)^T)V']O$

Proof. First Let X' be the output of f_2, which is the attention layer. Then we have:

$$\begin{aligned}
Rad^{abs}(X') &= Rad^{abs}(X) + Rad^{abs}\left(\text{softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)(XW_V)W_O\right) \\
&\leq Rad^{abs}(X) + Rad^{abs}\left(\text{softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)(XW_V)\right)O \\
&\leq Rad^{abs}(X) + [Rad^{abs}\left(\text{softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)\right)V \\
&+ \left\|\text{softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)\right\|_\infty Rad^{abs}(XW_V)]O \\
&\leq \left[\frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}Rad^{abs}(XW_Q(XW_K)^T) + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \mathbf{1}_{n\times n}Rad^{abs}(XW_V)\right]O \\
&\leq \left[\frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}\left(Rad^{abs}(XW_Q)K' + Q'Rad^{abs}(XW_K)^T\right) + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \mathbf{1}_{n\times n}Rad^{abs}(XW_V)\right]O \\
&\leq \left[\frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}\left(Rad^{abs}(X)QK' + Q'\left(Rad^{abs}(X)K\right)^T\right) + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \mathbf{1}_{n\times n}Rad^{abs}(X)V\right]O \\
&= f_1\left(Rad(X)\right)
\end{aligned} \tag{9}$$

While by previous result on Rademacher complexity of neural networks, immediately we have:

$$Rad(B) \leq |w|^T f_2(Rad(X'))$$

Combining two inequalities, our conclusion holds.

□

Theorem 8.

Let $h(X,Y,Z) = \text{softmax}\left(\frac{XY^T}{\sqrt{d_k}}\right)ZW_O$ where X, Y, Z are matrices, the multiplication is well-defined (softmax refers to row-wise softmax function), and d_k refers to the column dimension of X and Y.

Consider an L-layer Transformer network function class:

$$\mathcal{T}_L(X) = \left\{w^T\mathcal{B}_L \circ \mathcal{B}_{L-1} \circ \cdots \circ \mathcal{B}_1(X) + b\right\}$$

where each transformer block $\mathcal{B}_l(X)$ is defined as the composition of an MLP and an Attention layer:

$$\mathcal{B}_l(Z) = f_{l,\text{MLP}}(f_{l,\text{attn}}(Z))$$

with $f_{l,\text{attn}}(Z) = Z + h(ZW_{l,K}, ZW_{l,Q}, ZW_{l,V})$ and $f_{l,\text{MLP}}(Z) = Z + W_{l,2}g(W_{l,1}Z + B_{l,1}) + B_{l,2}$.

Assume the hypothesis space is restricted by the following matrix norm bounds for each layer $l \in [1, L]$: $|W_{l,2}| \leq \frac{2Y_{l,2}}{3}, |W_{l,1}| \leq \frac{2Y_{l,1}}{3}, |b_{l,2}| \leq \frac{2B_{l,2}}{3}, |b_{l,1}| \leq \frac{2B_{l,1}}{3}$, $|W_{l,Q}| \leq Q_l, |W_{l,K}| \leq K_l, |W_{l,V}| \leq V_l$, $|Z_{l-1}W_{l,Q}| \leq \frac{Q'_l}{2}, |Z_{l-1}W_{l,K}| \leq \frac{K'_l}{2}, |Z_{l-1}W_{l,V}| \leq \frac{V'_l}{2}$, $|W_{l,O}| \leq O_l$.

Then the empirical Rademacher complexity of the network is bounded by:

$$Rad(\mathcal{T}_L) \leq |w|^T\mathcal{M}_L \circ \mathcal{A}_L \circ \mathcal{M}_{L-1} \circ \mathcal{A}_{L-1} \cdots \circ \mathcal{M}_1 \circ \mathcal{A}_1\left(Rad^{abs}(X)\right)$$

where the complexity mapping operators for the MLP layer ($\mathcal{M}_l$) and the Attention layer ($\mathcal{A}_l$) are defined element-wise as:

For the MLP Layers (all $l \in [1, L]$):

$$\mathcal{M}_l(Z) = Z + Y_{l,2}\left(Y_{l,1}Z + \frac{B_{l,1}}{\sqrt{m}}\right) + \frac{B_{l,2}}{\sqrt{m}}$$

For the Attention Layers:

$$\mathcal{A}_l(Z) = Z + \left[\mathbf{1}_{n\times n} Z V_l + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}\left(ZQ_lK_l'^T + Q_l'(ZK_l)^T\right)V_l'\right]O_l$$

Proof. We prove the result by induction. Let X be the empirical input matrix. We prove the stronger statement that for the intermediate representations at each layer l, defined as the output of the attention layer Z_l and the output of the MLP layer H_l, the following element-wise inequalities hold:

$$Rad(H_l) \leq \mathcal{M}_l \circ \mathcal{A}_l \circ \cdots \circ \mathcal{M}_1 \circ \mathcal{A}_1\left(Rad^{abs}(X)\right)$$

For the first residual block ($l = 1$), the data first passes through the attention layer:

$$Z_1 = X + h(XW_{1,K}, XW_{1,Q}, XW_{1,V})$$

By the subadditivity of Rademacher complexity,

$$Rad(Z_1) \leq Rad(X) + Rad(h(XW_{1,K}, XW_{1,Q}, XW_{1,V}))$$

Using the structural bound for the attention mechanism derived previously, we obtain:

$$Rad(Z_1) \leq \mathcal{A}_1\left(Rad^{abs}(X)\right)$$

Next, Z_1 passes through the first MLP layer:

$$H_1 = Z_1 + W_{1,2}g(W_{1,1}Z_1 + B_{1,1}) + B_{1,2}$$

By subadditivity, Talagrand's contraction lemma (since g is 1-Lipschitz), and the scaling lemma (exactly as in the ResNet proof),

$$Rad(H_1) \leq Rad(Z_1) + Y_{1,2}\left(Y_{1,1}Rad(Z_1) + \frac{B_{1,1}}{\sqrt{m}}\right) + \frac{B_{1,2}}{\sqrt{m}}$$

This corresponds exactly to the definition of the MLP complexity operator $\mathcal{M}_1$. Thus,

$$Rad(H_1) \leq \mathcal{M}_1(Rad(Z_1)) \leq \mathcal{M}_1 \circ \mathcal{A}_1\left(Rad^{abs}(X)\right)$$

Hence, the claim holds for the first layer.

Now assume the statement holds for the k-th layer output H_k. Consider the $(k+1)$-th block. For the attention layer:

$$Z_{k+1} = H_k + h(H_kW_{k+1,K}, H_kW_{k+1,Q}, H_kW_{k+1,V})$$

By previous result of attention mechanism on Rademacher complexity , similarly we have:

$$Rad^{abs}(Z_{k+1}) \leq Rad^{abs}(H_k) + \Bigg[\mathbf{1}_{n\times n}Rad^{abs}(H_k)V_{k+1} + \frac{1}{n\sqrt{m}}\mathbf{1}_{n\times d_k} + \frac{1}{\sqrt{d_k}}\mathbf{1}_{n\times n}\left(Rad^{abs}(H_k)Q_{k+1}K_{k+1}'^T + Q_{k+1}'(Rad^{abs}(H_k)K_{k+1})^T\right)V_{k+1}'\Bigg]O_{k+1} \tag{10}$$

This matches the definition of the deep attention operator $\mathcal{A}_{k+1}$ which retains the residual term $Rad(H_k)$ (equivalent to $Z+$ in the operator definition). Thus,

$$Rad^{abs}(Z_{k+1}) \leq \mathcal{A}_{k+1}(Rad^{abs}(H_k))$$

Subsequently, for the $(k+1)$-th MLP layer:

$$H_{k+1} = Z_{k+1} + W_{k+1,2}g(W_{k+1,1}Z_{k+1} + B_{k+1,1}) + B_{k+1,2}$$

Applying subadditivity and the contraction lemma as before,

$$Rad^{abs}(H_{k+1}) \leq Rad^{abs}(Z_{k+1}) + Y_{k+1,2}\left(Y_{k+1,1}Rad^{abs}(Z_{k+1}) + \frac{B_{k+1,1}}{\sqrt{m}}\right) + \frac{B_{k+1,2}}{\sqrt{m}}$$

Which gives,

$$Rad^{abs}(H_{k+1}) \leq \mathcal{M}_{k+1}(Rad^{abs}(Z_{k+1}))$$

Combining the attention and MLP steps for layer $k+1$, we obtain:

$$Rad^{abs}(H_{k+1}) \leq \mathcal{M}_{k+1} \circ \mathcal{A}_{k+1}(Rad^{abs}(H_k))$$

By the induction hypothesis for $Rad(H_k)$,

$$Rad^{abs}(H_{k+1}) \leq \mathcal{M}_{k+1} \circ \mathcal{A}_{k+1} \circ \mathcal{M}_k \circ \mathcal{A}_k \circ \cdots \circ \mathcal{M}_1 \circ \mathcal{A}_1 \left(Rad^{abs}(X)\right)$$

Hence, the induction hypothesis is preserved for all L layers.

Finally, we consider the output projection of the Transformer class $\mathcal{T}_L(X) = w^T H_L$. By subadditivity and the property of scalar multiplication in Rademacher complexity,

$$Rad^{abs}(\mathcal{T}_L) \leq |w|^T Rad^{abs}(H_L)$$

Substituting the bounded result for $Rad(H_L)$ derived via induction yields the final theorem:

$$Rad^{abs}(\mathcal{T}_L) \leq |w|^T \mathcal{M}_L \circ \mathcal{A}_L \circ \cdots \circ \mathcal{M}_1 \circ \mathcal{A}_1 \left(Rad^{abs}(X)\right)$$

This completes the proof.

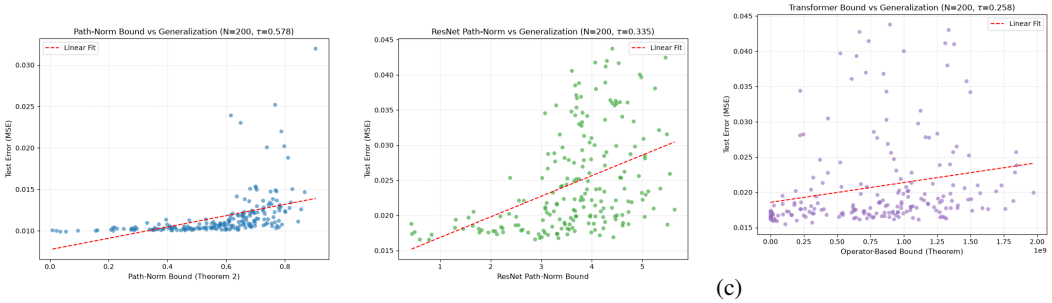
□

B Empirical Validation of Path Norm Complexity Bounds

To empirically substantiate our derived generalization bounds, we conduct systematic observational evaluations assessing the rank correlation between theoretical operator bounds and true generalization error (test MSE). Following standard evaluation protocols for generalization measures, we explore the hyperparameter landscape across diverse deep architectures to verify whether the operator-based capacity metrics reliably predict out-of-sample risk.

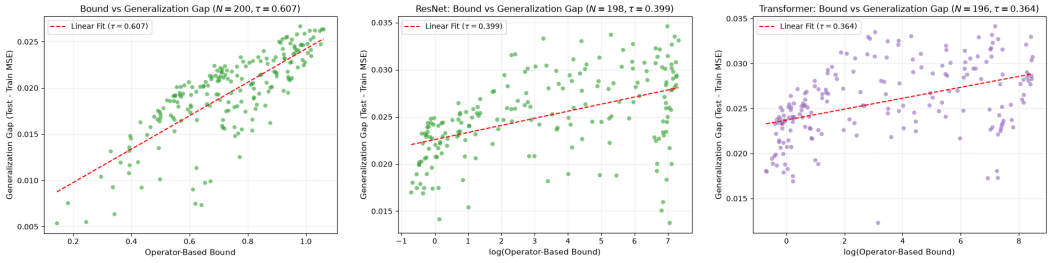
B.1 Experimental Setup

Teacher-Student Benchmark. We employ a controlled teacher-student synthetic regression framework across three canonical model classes: standard feedforward networks (MLP), residual networks (ResNet), and single-block self-attention networks (Transformer). For each architecture family, a fixed overparameterized teacher network with identical inductive structure is randomly initialized to generate target responses $y = f_{\text{teacher}}(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$ introduces bounded label noise. For the MLP and ResNet experiments, we set the teacher depth to 6 and the student depth to 2, while for the Transformer experiments, we use teacher and student depths of 4 and 2, respectively. For each experiment, the teacher network generates 2,000 randomly sampled training examples for the student. The width of all neural networks is fixed at 64 across all architectures.



(a) MLP ($N = 200, \tau = 0.578$) (b) ResNet ($N = 200, \tau = 0.335$) (c) Transformer ($N = 200, \tau = 0.258$)

Figure 1: Correlation between our theoretical operator-based bounds and generalization error (Test MSE) across $N = 200$ models. The red dashed line denotes the linear trend. Across all architectures, the theoretical bounds exhibit lower-bound envelope behavior at lower values and effectively capture generalization variance as the bound magnitude increases.



(a) MLP ($N = 200, \tau = 0.607$) (b) ResNet ($N = 198, \tau = 0.399$) (c) Transformer ($N = 196, \tau = 0.364$)

Figure 2: Empirical evaluation of operator-based complexity bounds against the **Generalization Gap** ($\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$) across diverse architectures. Red dashed lines denote linear trends. All architectures exhibit consistent rank correlation ($p < 10^{-10}$) and lower-bound enveloping.

B.2 Rank Correlation Analysis

We measure the alignment between our theoretical log-bound $\mathcal{B}(f)$ and test MSE using Kendall's rank correlation coefficient τ, alongside the corresponding two-sided p-value. Table 2 summarizes the statistical alignment across the three evaluated model classes.

Table 2: Empirical evaluation of operator-based bounds across $N = 200$ student models per architecture. All reported Kendall's τ values exhibit strong statistical significance ($p < 10^{-4}$).

Architecture	Activation / Layer	Sample Size (N)	Kendall's τ	p-value
MLP (Theorem 1)	GELU	200	**0.578**	$< 10^{-15}$
ResNet (Theorem 2)	Residual Blocks	200	**0.335**	$< 10^{-6}$
Transformer (Theorem 3)	Self-Attention + MLP	200	**0.258**	$< 10^{-4}$

B.3 Generalization Gap Evaluation

To rigorously assess the uniform convergence behavior governed by our theoretical framework, we evaluate the rank correlation between our derived bounds and the empirical generalization gap, defined as $\text{Gap}(f) = \mathcal{L}_{\text{test}}(f) - \mathcal{L}_{\text{train}}(f)$. As shown in Figure 2, subtracting empirical training error isolates over-fitting from optimization residuals, yielding stronger alignment with theoretical bounds across all architectures.

Table 3: Empirical rank correlation between operator-based complexity bounds and the Generalization Gap ($\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$) across architectures.

Architecture	**Evaluated Models** (N)	**Kendall's** τ	p**-value**
MLP	200	0.607	$< 10^{-18}$
ResNet	198	0.399	$< 10^{-16}$
Transformer	196	0.364	$< 10^{-12}$