Reproducing Kernel Banach Space Models for Neural Networks with Application to Rademacher Complexity Analysis

Alistair Shilton, Sunil Gupta, Santu Rana, Svetha Venkatesh

Applied Artificial Intelligence Initiative,
Deakin University, Geelong, Australia
alistair.shilton@deakin.edu.au, sunil.gupta@deakin.edu.au,
santu.rana@deakin.edu.au, svetha.venkatesh@deakin.edu.au

Abstract

This paper explores the use of Hermite transform based reproducing kernel Banach space methods to construct *exact* or *un-approximated* models of feedforward neural networks of arbitrary width, depth and topology, including ResNet and Transformers networks, assuming only a feedforward topology, finite energy activations and finite (spectral-) norm weights and biases. Using this model, two straightforward but surprisingly tight bounds on Rademacher complexity are derived, precisely (1) a general bound that is width-independent and scales exponentially with depth; and (2) a width- and depth-independent bound for networks with appropriately constrained (below threshold) weights and biases.

1 Introduction

A significant challenge in neural networks is understanding how large models, despite their high capacity to overfit training data, can still generalize effectively (Neyshabur et al., 2014). Learning theory tells us that inductive bias plays an important role in explaining this phenomena, where inductive bias is the restriction of the space of potential learned functions (neural networks) to a small subset $\mathcal{F}$ of the total space of space, either explicitly through regularization or implicitly through the training algorithm used. Rademacher complexity (Bartlett & Mendelson, 2002; Steinwart & Christman, 2008) is one measure of the complexity or expressive power of $\mathcal{F}$ that has been used to understand inductive bias through the lens of uniform convergence - that is, the rate at which the empirical risk (on the training dataset of N samples) converges to actual risk (on the data distribution) (for a discussion of alternative approaches see (Valle-Pérez & Louis, 2020)). A representative approach to Rademacher complexity analysis in neural networks is "peeling" (Neyshabur et al., 2015; Golowich et al., 2018; Truong, 2022). In this approach, the total compexity is bounded by "peeling off" the output layer D to extract factors due to that layer and thus express the total Racehmacher complexity as a product of terms due to the output layer D and the Rademacher complexity of the preceeding $(D-1)$-layer network. The process is then repeated, peeling off successive layers until the process terminates at the input layer. This typically results in a bound that exhibits width independence (assuming popular schemes such as LeCun, He or Glorot weight scaling) and exponential depth dependence, and contains some (typically Lipschitz-type) scaling term due to the neural activations as well a (typically depth-exponential) "nuisance factor". For example in (Neyshabur et al., 2015) it is shown that, for a simple unbiased layerwise network with spectral-norm bound weight matrices $\mathbf{W}^{[j-1:j]}$ and Lipschitz activations, the Rademacher complexity is bounded as:

$$\mathcal{R}_N\left(\mathcal{F} : \left\|\mathbf{W}^{[j-1:j]}\right\| \leq \omega^{[j-1:j]}\right) \sim \mathcal{O}\left(\frac{2^D \prod_{j=1}^D \omega^{[j-1:j]}}{\sqrt{N}}\right)$$

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

This bound can be refined in various ways (eg. (Golowich et al., 2018)), but the basic form remains, as do the nuisance factors (the term 2^D in the above bound, for example) in one form or another.

An alternative approach is to construct a bilinear (dual) representation of the model that splits the input $\boldsymbol{x} \in \mathbb{X}$ and parameters $\Theta \in \mathbb{W}$ into separate terms in a dual representation:

$$\mathbf{f}(\boldsymbol{x}; \Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\boldsymbol{x})]$$

where $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W}$, $\boldsymbol{\phi} : \mathbb{X} \to \mathcal{X}$ are feature maps and $\langle \cdot, \cdot] : \mathcal{W} \times \mathcal{X} \to \mathcal{R}$ is a continuous bilinear product. Examples of this type of model are the neural network Gaussian process (Rasmussen & Williams, 2006) (NNGP) models (Neal, 1996), which treat all layers prior to the output as fixed and model the influence of the weights in the output layer; neural tangent kernel (NTK) models (Jacot et al., 2018; Daniely, 2017; Daniely et al., 2016), which model the (first-order) variation of the weights about their initial values in a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) (see for example (Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2020; Zou & Gu, 2019; Arora et al., 2019b,a; Cao & Gu, 2019)); and reproducing kernel Banach space (RKBS) (Lin et al., 2022; Zhang et al., 2009; Zhang & Zhang, 2012; Song et al., 2013; Sriperumbudur et al., 2011; Xu & Ye, 2014) approaches such as (Shilton et al., 2023), which recursively construct feature maps $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W}$, $\boldsymbol{\phi} : \mathbb{X} \to \mathcal{X}$ to exactly model the neural network (beyond first-order).[1] In all cases the utility of the model in the context of Rademacher complexity analysis is that it makes the construction of bounds straightforward through the use of either the Cauchy-Schwarz inequality (if $\langle \cdot, \cdot]$ is an inner product) or the continuity of the bilinear product; and moreover, as peeling is not applied directly to the Rademacher complexity, nuisance factors arising from this procedure may be avoided. However the assumptions made by these models (wide-networks, lazy training, restrictions on neural activations and network topology etc (Arora et al., 2019b; Lee et al., 2019; Bai & Lee, 2019)) can complicate analysis and limit their applicability.

Our goal in this paper is to address two questions, (1) can we formulate an *exact* (non-approximate) model for a wide class of neural networks, including ResNet and Transformers, avoiding entirely the question of gaps between the performance of the neural network and its model; and (2) can such a model be used to derive straightforward, non-vacuous, widely applicable, training-independent bounds on Rademacher complexity *without* nuisance factors. We answer these questions with the following contributions:

1. **Exact RKBS model (Theorem 1):** For feedforward neural network with arbitrary topology, finite weight and biases and finite-energy neural activations, we construct an exact model that recasts neural networks as elements in a reproducing kernel Banach space (RKBS) defined by the bilinear product:

   $$\mathbf{f}(\boldsymbol{x}; \Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\boldsymbol{x})]_{\mathbf{g}}$$

   where $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W}$ is a weight/bias feature map, $\boldsymbol{\phi} : \mathbb{X} \to \mathcal{X}$ is a data feature map, and $\langle \cdot, \cdot]_{\mathbf{g}} : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ is a continuous bilinear form characterized by an indefinite metric $\mathbf{g}$.
2. **Rademacher Complexity Bound (Theorem 4):** We observe that, for our RKBS model:

   $$\|\mathbf{f}(\boldsymbol{x}; \Theta)\|_2 \leq C_\Theta \|\boldsymbol{\phi}(\boldsymbol{x})\|_2$$

   where $C_\Theta \leq 1$ and, using this, derive a straightforward non-asymptotic bound for the Rademacher complexity of a very general class of neural networks (including ResNet and Transformers) that is width-independent, depth-exponential and contains no nuisance-factors. For example for a scalar-valued, layerwise, fully-connected, unbiased ReLU network of depth D, our bound is exactly:

   $$\mathcal{R}_N\left(\mathcal{F} : \|\mathbf{W}^{[j-1:j]}\|_2 \leq \omega^{[j-1:j]}\right) \leq \frac{\prod_{j=1}^{D} \omega^{[j-1:j]}}{\sqrt{N}}$$

   where $\|\cdot\|_2$ is the spectral norm. More generally, we derive conditions under which the Rademacher complexity bound is both *width-* and *depth-independent*, and subsequently $\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$, and discuss implications for ReLU, ResNet and Transformer networks.

[1] Beyond bilinear RKBS (Lin et al., 2022), more general RKBS models have been used in eg. (Bartolucci et al., 2023; Sanders, 2020; Shilton et al., 2023; Parhi & Nowak, 2021; Unser, 2021, 2019; Spek et al., 2022).

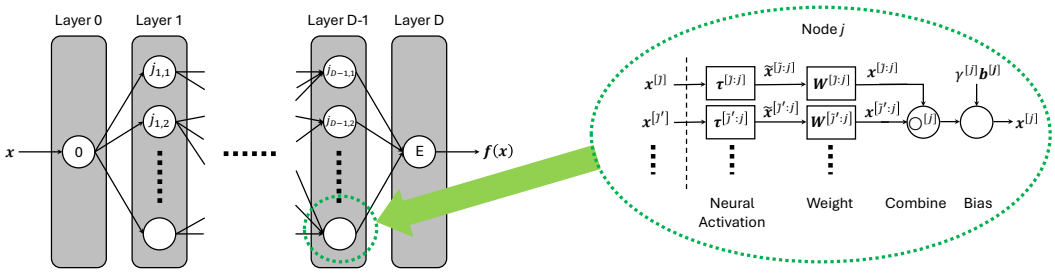


Figure 1: Layerwise feedforward neural network structure. Each layer $\boldsymbol{\jmath} \in \mathbb{Z}_D$ contains nodes $\mathbb{L}^{[\boldsymbol{\jmath}]}$, where the output of the node is computed as shown in the inset. Note that a computational skeleton (Daniely et al., 2016) with one input and one output can be modified to this form by inserting skip nodes (nodes with $\mathbb{A}^{[j]} = \{\widetilde{\jmath}\}$, $\mathbf{W}^{[\widetilde{\jmath}:j]} = \mathbf{I}$, $\mathbf{b}^{[j]} = \mathbf{0}$, $\boldsymbol{\tau}^{[\widetilde{\jmath}:j]} = \mathrm{id}$) into the graph as required.

1.1 Mathematical Notations

Vectors and matrices: Column vectors are $\mathbf{a}, \mathbf{b}$ with elements a_i, b_j. Matrices are $\mathbf{A}, \mathbf{B}$ with elements $A_{i,j}$, rows $\mathbf{A}_{i:}$ and columns $\mathbf{A}_{:j}$. $|\mathbf{a}|$ and $\mathrm{sgn}(\mathbf{a})$ are the elementwise norm and sign. $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$ is the spectral norm and $\|\mathbf{A}\|_F$ is the Frobenius norm. $\mathbf{A} \odot \mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$, $\mathbf{A} \otimes^{\updownarrow} \mathbf{B}$ are Hadamard, Kronecker and columnwise Kronecker (Khatri-Rao) product. $\mathbf{A} \oplus \mathbf{B} = [\mathbf{A}^{\mathrm{T}}\ \mathbf{B}^{\mathrm{T}}]^{\mathrm{T}}$ is columnar matrix concatenation. $\mathbf{A}^{\bigcirc k} = \mathbf{A} \bigcirc^{k \text{ times}} \cdots \bigcirc \mathbf{A}$ is the exponentiation for operator $\bigcirc$.

Products and norms: $\langle \cdot, \cdot \rangle$, $\langle\!\langle \cdot, \cdot, \cdot, \ldots \rangle\!\rangle$ and $\langle \cdot, \cdot]$ are inner, multilinear and bilinear products, where $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_i a_i b_i$, $\langle\!\langle \mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots \rangle\!\rangle = \sum_i a_i b_i c_i \ldots$, $\langle \mathbf{a}, \mathbf{b}]_{\mathbf{g}} = \sum_i g_i a_i b_i$ and $\langle \mathbf{A}, \mathbf{b}]_{\mathbf{g}} = \sum_i g_i \mathbf{A}_{i:} b_i$ throughout. We also find it convenient to define an operator form $\langle\!\langle \cdot \rangle\!\rangle_i \mathbf{a}_i = \langle\!\langle \mathbf{a}_1, \mathbf{a}_2, \ldots \rangle\!\rangle$.

Sets and functions: $\mathbb{N} = \{0, 1, \ldots\}$, $\mathbb{Z}_+ = \{1, 2, \ldots\}$, $\mathbb{Z}_N = \{1, \ldots, N\}$. $\partial\mathbb{A}$ is the boundary of $\mathbb{A}$. $\mathrm{id}(\mathbf{a}) = \mathbf{a}$. $[a]_+ = \max\{a, 0\}$, $[\mathbf{a}]_+ = [[a_i]_+]_i$. $\mathbf{a}^{\langle \odot b \rangle} = \mathrm{sgn}(\mathbf{a}) \odot |\mathbf{a}|^{\odot b}$ (Der & Lee, 2007). $L^2(\mathbb{R}^{\tilde{H}}, e^{-\|\boldsymbol{\zeta}\|_2^2}) = \{\boldsymbol{\tau} : \mathbb{R}^{\tilde{H}} \to \mathbb{R}^H \,|\, \int_{\boldsymbol{\zeta} \in \mathbb{R}^{\tilde{H}}} \|\boldsymbol{\tau}(\boldsymbol{\zeta})\|_2^2 e^{-\|\boldsymbol{\zeta}\|_2^2} d\boldsymbol{\zeta} < \infty\}$ are the finite-energy functions.

Multi-indices: Multi-indices are $\mathtt{k}, \mathtt{l} \in \mathbb{N}^n$ with elements k_i, l_j. $|\mathtt{k}| = \sum_i k_i$, $\mathtt{k}! = \prod_i k_i!$, $\mathbf{a}^{\mathtt{k}} = \prod_i a_i^{k_i}$, $\binom{\mathtt{k}}{\mathtt{l}} = \prod_i \binom{k_i}{l_i}$. $\frac{\partial^{\mathtt{k}}}{\partial \boldsymbol{x}^{\mathtt{k}}} = \prod_i \frac{\partial^{k_i}}{\partial x_i^{k_i}}$. We use the shorthands $\mathtt{k} \succ_n l$ for $\mathtt{k} \in \mathbb{N}^n$ and $|\mathtt{k}| > l$, $\mathtt{k} \succeq_n l$ for $\mathtt{k} \in \mathbb{N}^n$ and $|\mathtt{k}| \geq l$, $\mathtt{k} \prec_n l$ for $\mathtt{k} \in \mathbb{N}^n$ and $|\mathtt{k}| < l$, $\mathtt{k} \preceq_n l$ for $\mathtt{k} \in \mathbb{N}^n$ and $|\mathtt{k}| \leq l$.

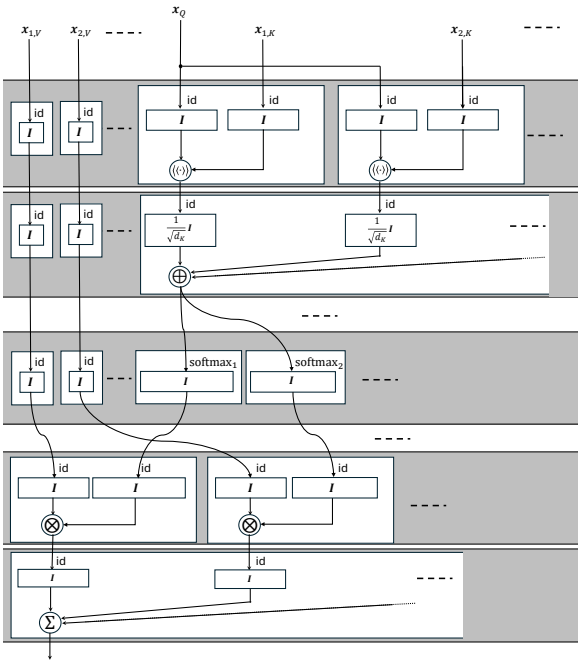
Hermite Polynomials: $He_k(x)$ are the (probabilist's) Hermite polynomials. $He_{\mathtt{k}}(\boldsymbol{x}) = \prod_i He_{k_i}(x_i)$ are the multivariate Hermite polynomials. $\mathrm{He}_k = He_k(0)$, $\mathrm{He}_{\mathtt{k}} = He_{\mathtt{k}}(\mathbf{0})$ are the Hermite numbers (Abramowitz et al., 1972; Morse & Feshbach, 1953; Olver et al., 2010; Rahman, 2017).

Indexing Conventions: Layers are $\boldsymbol{\jmath} \in \mathbb{Z}_D$ (there are D layers). Nodes are $j \in \mathbb{Z}_E$ (there are E nodes). Layer $\boldsymbol{\jmath}$ contains nodes $\mathbb{L}^{[\boldsymbol{\jmath}]} \subseteq \mathbb{Z}_E$: $\cup_{\boldsymbol{\jmath} \in \mathbb{Z}_D} \mathbb{L}^{[\boldsymbol{\jmath}]} = \mathbb{Z}_E$, $\mathbb{L}^{[\boldsymbol{\jmath}]} \cap \mathbb{L}^{[\boldsymbol{\jmath}']} = \emptyset \forall \boldsymbol{\jmath} \neq \boldsymbol{\jmath}'$. Node $j \in \mathbb{L}^{[\boldsymbol{\jmath}]}$ in layer $\boldsymbol{\jmath}$ has parents $\widetilde{\jmath} \in \mathbb{A}^{[j]} \subseteq \mathbb{L}^{[\boldsymbol{\jmath}-1]}$. $\mathbb{L}^{[0]} = \{0\}$, $\mathbb{L}^{[D]} = \{E\}$ are the input/output layers.

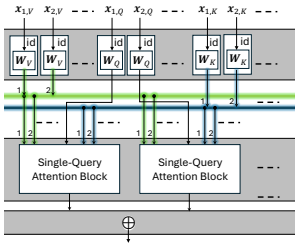
2 Setting and Assumptions

We consider layerwise feedforward neural networks as shown in Figure 1. This contains E nodes $j \in \mathbb{Z}_E$ arranged in D layers $\boldsymbol{\jmath} \in \mathbb{Z}_D$ and a virtual input node $j = 0$ (in virtual layer $\boldsymbol{\jmath} = 0$), where layer $\boldsymbol{\jmath} \in \mathbb{Z}_D$ contains nodes $\mathbb{L}^{[\boldsymbol{\jmath}]} \subseteq \mathbb{Z}_E$ and layer D contains a single output node E. A node $j \in \mathbb{L}^{[\boldsymbol{\jmath}]}$ has parents $\mathbb{A}^{[j]} \subseteq \mathbb{L}^{[\boldsymbol{\jmath}-1]}$, with its function being specified by an operator $\bigcirc^{[j]} \in \{\oplus, \sum, \otimes, \langle\!\langle \cdot \rangle\!\rangle\}$. Given input $\boldsymbol{x}$, data flows from node $j = 0$ to node $j = E$ as per Figure 1:

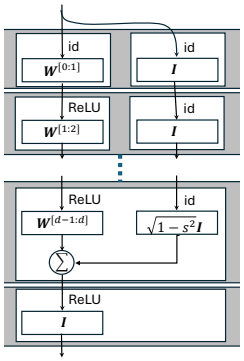
$$\left.\begin{array}{c} \boldsymbol{x}^{[0]} = \boldsymbol{x} \\ \downarrow \\ \left.\begin{array}{l} \boldsymbol{x}^{[j]} = \bigcirc_{\widetilde{\jmath} \in \mathbb{A}^{[j]}}^{[j]} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]} + \gamma^{[j]} \mathbf{b}^{[j]} \in \mathbb{R}^{H^{[j]}} \\ \boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\boldsymbol{x}^{[\widetilde{\jmath}]}) \in \mathbb{R}^{H^{[\widetilde{\jmath}:j]}} \quad \forall \widetilde{\jmath} \in \mathbb{A}^{[j]} \end{array}\right] \forall j \in \mathbb{Z}_E \quad \text{(a)} \\ \downarrow \\ \mathbf{f}(\boldsymbol{x}; \Theta) = \boldsymbol{x}^{[E]} \end{array}\right\} \tag{1}$$



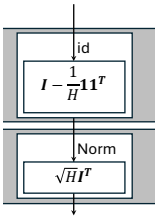
(a) Single-Query Attention Block



(b) Attention Block



(c) Residual Block



(d) LayerNorm Block

Block Type	Norm-bound (see Appendices D.1, D.2, D.3 and D.4)	
Residual Block	$\phi^2 = \left(1 - s^2 + \prod_{q-1}^{d} \omega^{[q-1:q]2}\right) \phi_{\text{in}}^2$	
LayerNorm Block	$\phi^2 = H$	
Single-Query Attention Block	$\phi^2 = \lambda^2 \phi_V^2 \phi_Q^2 \phi_K^2$	
Single-Head Attention Block	$\phi^2 = q\lambda^2 \omega_V^2 \omega_Q^2 \omega_K^2 \phi_V^2 \phi_Q^2 \phi_K^2$	(for q queries)

(e) Bounds Relevant to Section 2.1

Figure 2: Residual, attention and LayerNorm blocks. In the residual block $s \in (0, 1)$. The single-query attention block (a) is for a single query $\boldsymbol{x}_{1,Q}$ with keys $\boldsymbol{x}_{1,K}$, $\boldsymbol{x}_{2,K}$, ... and values $\boldsymbol{x}_{1,V}$, $\boldsymbol{x}_{2,V}$, ... A single-head attention block (b) is formed from multiple single-query blocks (the usual (matrix) output has been vectorized here). See Table 3 for definitions of neural activations used here.

where $\boldsymbol{\tau}^{[\widetilde{j}:j]} : \mathbb{R}^{H^{[\widetilde{j}]}} \to \mathbb{R}^{H^{[\widetilde{j}:j]}}$ are neural activation functions; $\mathbf{W}^{[\widetilde{j}:j]}$ are weight matrices; $\mathbf{b}^{[j]}$ biases; $\gamma^{[j]} \in \{0, 1\}$ (unbiased and biased); and $\Theta = \{\mathbf{W}^{[\widetilde{j}:j]}, \mathbf{b}^{[j]} : j \in \mathbb{Z}_E, \widetilde{j} \in \mathbb{A}^{[j]}\}$. We assume that:

1. **Bounded inputs:** $\boldsymbol{x} \in \mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq 1\}$.
2. **Finite weights/biases:** $\Theta \in \mathbb{W} = \{\mathbf{W}^{[\widetilde{j}:j]}, \mathbf{b}^{[j]} : \|\mathbf{W}^{[\widetilde{j}:j]}\|_2, \|\mathbf{b}^{[j]}\|_2 < \infty : \widetilde{j} \in \mathbb{A}^{[j]}, j \in \mathbb{Z}_E\}$.
3. **Finite activations:** $\boldsymbol{\tau}^{[\widetilde{j}:j]} \in L^2(\mathbb{R}^{H^{[\widetilde{j}]}}, e^{-\|\boldsymbol{\varsigma}\|_2^2}) \; \forall \widetilde{j} \in \mathbb{A}^{[j]}, j \in \mathbb{Z}_E$.
4. **Lipschitz/Bounded activations:** $\boldsymbol{\tau}^{[\widetilde{j}:j]}$ is either Lipschitz or bounded $\forall \widetilde{j} \in \mathbb{A}^{[j]}, j \in \mathbb{Z}_E$, and $\boldsymbol{x} \in \partial\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$ if any $\boldsymbol{\tau}^{[\widetilde{j}:j]}$ are non-Lipschitz.)

Note that assumption 4 is not required when constructing our bilinear feature space model of neural networks, but is required to cast this model in RKBS and subsequently derive our Rademacher complexity bound. The set of neural networks satisfying our assumptions is $\mathcal{F}$, and its dual is $\mathcal{F}^\star$:

$$\begin{aligned} \mathcal{F} &= \{\mathbf{f}(\cdot; \Theta) : \mathbb{X} \to \mathbb{R}^m | \; \mathbf{f} \text{ as per (1) satisfying assumptions 1-4}\} \\ \mathcal{F}^\star &= \{\mathbf{f}(\boldsymbol{x}; \cdot) : \mathbb{W} \to \mathbb{R}^m | \; \mathbf{f} \text{ as per (1) satisfying assumptions 1-4}\} \end{aligned} \tag{2}$$

This model is rather general to encompass a wider variety of network architectures. Residual (He et al., 2016) blocks can be built using additive nodes $\bigcirc^{[j]} = \sum$ as shown in Figure 2c. Single

Neural Activation		$\boldsymbol{\tau} \in L^2(\mathbb{R}^H, e^{-\|\boldsymbol{\zeta}\|_2^2})$	Lipschitz (L_r)	Bounded (B)	Valid here
Linear	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = \boldsymbol{\zeta}$	✓	✓ (1)	×	✓
ReLU	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = [\boldsymbol{\zeta}]_+$	✓	✓ (1)	×	✓
Poly-ReLU	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = [\boldsymbol{\zeta}]_+^{\odot p}$	✓	✓ (pa^{p-1})	×	✓
Tanh	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = [\tanh(\zeta_i)]_i$	✓	✓ (1)	-	✓
Sigmoid	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = \left[\frac{1}{1+e^{-\zeta_i}}\right]$	✓	✓ ($\frac{1}{2}$)	-	✓
Softmax	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = \left[\frac{e^{\lambda\zeta_{i'}}}{\sum_{i''} e^{\lambda\zeta_{i''}}}\right]_{i'}$	✓	✓ (λ)	-	✓
Softmax$_i$	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = \left[\delta_{i,i'}\frac{e^{\lambda\zeta_{i'}}}{\sum_{i''} e^{\lambda\zeta_{i''}}}\right]_{i'}$	✓	✓ (λ)	-	✓
Norm	$\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) = \frac{\boldsymbol{\zeta}}{\|\boldsymbol{\zeta}\|_2}$	✓	×	✓ (1)	✓

Figure 3: Characteristics of common neural activation functions. We include poly-ReLU (Cho & Saul, 2009) here as an example where the Lipschitz constant L_a of $\boldsymbol{\tau}|_a$ depends on the radius a. See (Gao & Pavel, 2017) for more detail regarding the softmax Lipschitz constant.

query attention blocks (Vaswani et al., 2017) can be constructed as shown in Figure 2a using not just additive but also inner-product $\bigcirc^{[j]} = \langle\!\langle \cdot \rangle\!\rangle$, multiplicative $\bigcirc^{[j]} = \otimes$ and columnar concatenation $\bigcirc^{[j]} = \oplus$ nodes. Full attention block can be constructed as shown in Figure 2b (and similarly multi-head attention). Finally, a LayerNorm (layer normalization (Ba et al., 2016)) block is shown in Figure 2d. We note that blocks of this sort may be combined to form more general networks. Later, we find it convenient to include non-trivial nodes or blocks in the network, so for example we may speak of an "attention node" j that encompasses (abstracts away) a complete attention block (Figure 2b).

2.1 Characterization of Neural Activations

As noted previously, we assume all activation functions in the network are Lipschitz/bounded and finite energy. The finite-energy assumption allows us to apply the Hermite transform to the neural activation functions and subsequently construct our bilinear model of the network, while the Lipschitz/bounded property suffices to ensure that the bilinear model is continuous. Starting with the finite-energy assumption, the multivariate (probabilist's) Hermite polynomials (Abramowitz et al., 1972; Morse & Feshbach, 1953; Olver et al., 2010; Rahman, 2017) are, for multi-index $\mathtt{k} \in \mathbb{N}^n$:

$$He_{\mathtt{k}}(\boldsymbol{\zeta}) = (-1)^{|\mathtt{k}|} e^{\frac{1}{2}\|\boldsymbol{\zeta}\|_2^2} \frac{\partial^{\mathtt{k}}}{\partial \boldsymbol{\zeta}^{\mathtt{k}}} e^{-\frac{1}{2}\|\boldsymbol{\zeta}\|_2^2} = \sum_{0 \preceq_n \mathtt{l} \leq \mathtt{k}} \binom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \boldsymbol{\zeta}^{\mathtt{l}}$$

where $\mathrm{He}_{\mathtt{k}} = He_{\mathtt{k}}(\mathbf{0})$ are Hermite numbers. These form an orthogonal basis of $L^2(\mathbb{R}^n, e^{-\|\boldsymbol{\zeta}\|_2^2})$ (Appendix A). By assumption $\boldsymbol{\tau}^{[\tilde{\jmath}:j]} \in L^2(\mathbb{R}^{H^{[\tilde{\jmath}]}}, e^{-\|\boldsymbol{\zeta}\|_2^2})$, and thus the Hermite transform exists:[2]

$$\begin{aligned} \boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) &= \boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\mathbf{0}) + \sum_{\mathtt{k} \succ_{H^{[\tilde{\jmath}]}} 0} \mathbf{a}_{\mathtt{k}}^{[\tilde{\jmath}:j]} \sum_{0 \prec_{H^{[\tilde{\jmath}]}} \mathtt{l} \leq \mathtt{k}} \binom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \boldsymbol{\zeta}^{\mathtt{l}} \\ \text{where: } \mathbf{a}_{\mathtt{k}}^{[\tilde{\jmath}:j]} &= \tfrac{1}{\sqrt{2\pi \mathtt{k}!}} \int_{\boldsymbol{\zeta} \in \mathbb{R}^{H^{[\tilde{\jmath}]}}} \left(\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta}) - \boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\mathbf{0})\right) He_{\mathtt{k}}(\boldsymbol{\zeta}) e^{-\frac{1}{2}\|\boldsymbol{\zeta}\|_2^2} d\boldsymbol{\zeta} \end{aligned} \tag{3}$$

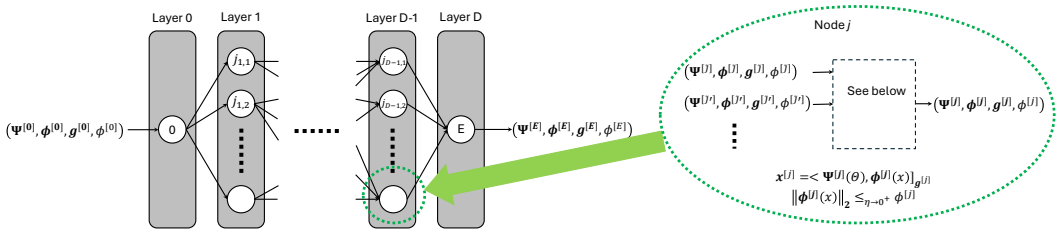
From this we define the magnitude functions $s_\eta^{[\tilde{\jmath}:j]} : \mathbb{R}_+ \to \mathbb{R}_+$ (where $\eta \in \mathbb{R}_+$):

$$s_\eta^{[\tilde{\jmath}:j]}(\zeta) = \eta^2 \left\|\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \sum_{\mathtt{k} \succ_{H^{[\tilde{\jmath}]}} 0} \left\|\mathbf{a}_{\mathtt{k}}^{[\tilde{\jmath}:j]}\right\|_1 \left((1+\eta\zeta)^{|\mathtt{k}|} - 1\right) \tag{4}$$

which are monotonically increasing and superadditive on $\mathbb{R}_+$. Note that while the Hermite transform terms and magnitude functions play an important role in the construction of our model, they play no role in our subsequent analysis of Rademacher complexity (they vanish in our analysis in the limit $\eta \to 0^+$). Thus, for our purposes, beyond their existence (which is guaranteed), their exact value/form does not matter here. Nevertheless, see Appendix B for a full analysis of the ReLU activation.

Regarding assumptions 4, if $\boldsymbol{\tau}^{[\tilde{\jmath}:j]}|_a$ ($\boldsymbol{\tau}^{[\tilde{\jmath}:j]}$ restricted to a ball of radius a) is Lipschitz then we denote the Lipschitz constant by $L_a^{[\tilde{\jmath}:j]}$. Conversely, if $\boldsymbol{\tau}^{[\tilde{\jmath}:j]}$ is absolutely bounded, we denote the bound $B^{[\tilde{\jmath}:j]}$, where $|\boldsymbol{\tau}^{[\tilde{\jmath}:j]}(\boldsymbol{\zeta})| \leq B^{[\tilde{\jmath}:j]}\ \forall \boldsymbol{\zeta}$. While assumption 4 is not required to construct our bilinear dual representation we find it useful to include $L_a^{[\tilde{\jmath}:j]}$ here to simplify later results. When $L_a^{[\tilde{\jmath}:j]}$ is ill-defined we use the nominal value $L_a^{[\tilde{\jmath}:j]} = B^{[\tilde{\jmath}:j]}/\phi^{[\tilde{\jmath}]}$ in the bounded case, or $L_a^{[\tilde{\jmath}:j]} = 1$ if assumption 4 is not satisfied. Table 3 provides Lipschitz constants/bounds for common neural activations.

[2]These are conditionally convergent series in general, so ordering of multi-indices $\mathtt{k}$, $\mathtt{l}$ in sums and vectors must be enforced consistently and must be compatible with the semi-ordering imposed by $\preceq_{H^{[\tilde{\jmath}]}}$.



$$
\left.
\begin{aligned}
&\boldsymbol{\Psi}^{[0]}(\Theta)=\mathbf{I}_n, \boldsymbol{\phi}^{[0]}(\boldsymbol{x})=\boldsymbol{x}, \mathbf{g}^{[0]}=\mathbf{1}_n, \psi^{[0]}=1, \phi^{[0]}=1 \\
&\qquad\qquad\downarrow \\
&\boldsymbol{\Psi}^{[j]}(\Theta)=\left[\begin{array}{c} \mathbf{b}^{[j]\mathrm{T}} \\ \square^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}} \frac{L_{\psi^{[j]}}^{[\widetilde{j}:j]}}{\sqrt{s_\eta^{[\widetilde{j}:j]}(1)}} \left[\begin{array}{c} \eta \mathrm{diag}\left(\boldsymbol{\tau}^{[\widetilde{j}:j]}(\mathbf{0})^{\langle\odot\frac{1}{2}\rangle}\right) \\ \left[\mathrm{diag}\left(\mathbf{a}_{\mathbf{k}}^{[\widetilde{j}:j]\langle\odot\frac{1}{2}\rangle}\right)\otimes^{\updownarrow}\left[\binom{\mathbf{k}}{\mathbf{1}}^{\frac{1}{2}}\bigotimes_{\widetilde{i}\in\mathbb{Z}_{H^{[\widetilde{j}]}}}\left(\sqrt{\eta}\boldsymbol{\Psi}_{:\widetilde{i}}^{[\widetilde{j}]}(\Theta)\right)^{\otimes l_{\widetilde{i}}}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0} \end{array}\right]\mathbf{W}^{[\widetilde{j}:j]} \end{array}\right] \\
&\boldsymbol{\phi}^{[j]}(\boldsymbol{x})=\left[\begin{array}{c} \beta^{[j]} \\ \square^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}} \frac{L_{\phi^{[j]}}^{[\widetilde{j}:j]}}{\sqrt{s_\eta^{[\widetilde{j}:j]}(1)}} \left[\begin{array}{c} \eta\left|\boldsymbol{\tau}^{[\widetilde{j}:j]}(\mathbf{0})\right|^{\odot\frac{1}{2}} \\ \left[\left|\mathbf{a}_{\mathbf{k}}^{[\widetilde{j}:j]}\right|^{\odot\frac{1}{2}}\otimes\left[\binom{\mathbf{k}}{\mathbf{1}}^{\frac{1}{2}}\left(\sqrt{\eta}\boldsymbol{\phi}^{[\widetilde{j}]}(\boldsymbol{x})\right)^{\otimes|\mathbf{1}|}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0} \end{array}\right]\omega^{[\widetilde{j}:j]} \end{array}\right] \\
&\mathbf{g}^{[j]}=\left[\begin{array}{c} \frac{\gamma^{[j]}}{\beta^{[j]}} \\ \square^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}} \frac{s_\eta^{[\widetilde{j}:j]}(1)}{L_{\psi^{[\widetilde{j}]}}^{[\widetilde{j}:j]}L_{\phi^{[\widetilde{j}]}}^{[\widetilde{j}:j]}} \left[\begin{array}{c} \frac{1}{\eta^2}\mathbf{1}_{H^{[\widetilde{j}:j]}} \\ \left[\mathbf{1}_{H^{[\widetilde{j}:j]}}\otimes\left[\mathrm{He}_{\mathbf{k}-\mathbf{1}}\left(\frac{1}{\eta}\mathbf{g}^{[\widetilde{j}]}\right)^{\otimes|\mathbf{1}|}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0} \end{array}\right]^{\frac{1}{\omega^{[\widetilde{j}:j]}}} \end{array}\right] \\
&\psi^{[j]2}=\beta^{[j]2}+\bigcirc^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}} L_{\psi^{[\widetilde{j}]}}^{[\widetilde{j}:j]2}\omega^{[\widetilde{j}:j]2}\psi^{[\widetilde{j}]2}, \quad \phi^{[j]2}=\beta^{[j]2}+\bigcirc^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}} L_{\phi^{[\widetilde{j}]}}^{[\widetilde{j}:j]2}\omega^{[\widetilde{j}:j]2}\phi^{[\widetilde{j}]2} \\
&\qquad\qquad\downarrow \\
&\boldsymbol{\Psi}(\Theta)=\boldsymbol{\Psi}^{[E]}(\Theta), \boldsymbol{\phi}(\boldsymbol{x})=\boldsymbol{\phi}^{[E]}(\boldsymbol{x}), \mathbf{g}=\mathbf{g}^{[E]}, \psi=\psi^{[E]}, \phi=\phi^{[E]}
\end{aligned}
\right\} \forall j\in\mathbb{Z}_E \quad \text{(a)} \qquad (5)
$$

Where: $\square^{[j]}=\begin{cases}\mathrm{diag} & \text{if } \bigcirc^{[j]}=\bigoplus \\ \bigoplus & \text{if } \bigcirc^{[j]}=\sum \\ \bigotimes & \text{if } \bigcirc^{[j]}=\bigotimes \\ (\bigotimes^{\updownarrow}\cdot)\mathbf{1} & \text{if } \bigcirc^{[j]}=\langle\!\langle\cdot\rangle\!\rangle\end{cases}$, $\boxdot^{[j]}=\begin{cases}\bigoplus & \text{if } \bigcirc^{[j]}\in\{\bigoplus,\sum\} \\ \bigotimes & \text{if } \bigcirc^{[j]}\in\{\bigotimes,\langle\!\langle\cdot\rangle\!\rangle\}\end{cases}$ and $\circledcirc^{[j]}=\begin{cases}\sum & \text{if } \bigcirc^{[j]}\in\{\bigoplus,\sum\} \\ \prod & \text{if } \bigcirc^{[j]}\in\{\bigotimes,\langle\!\langle\cdot\rangle\!\rangle\}\end{cases}$

Figure 4: Recursive definition of the bilinear representation $\mathbf{f}(\boldsymbol{x};\Theta)=\langle\boldsymbol{\Psi}(\Theta),\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathbf{g}}$. The upper figure is a schematic representation of the formal definition (5), where the bilinear representation of the output of each node is obtained, using (5a), from the bilinear representations of the inputs $\widetilde{j}\in\mathbb{A}^{[j]}$ to that node. Subsequently the bilinear representation of the network is defined recursively in terms of the (trivial) bilinear representation of $\boldsymbol{x}=\langle\mathbf{I},\boldsymbol{x}\rangle_{\mathbf{1}}$. $\eta\in\mathbb{R}_+$ is an (arbitrary) constant that will be helpful in our Rademacher complexity analysis. With regard notation, we recall that node j is characterized by its operation $\bigcirc^{[j]}\in\{\oplus,\sum,\otimes,\langle\!\langle\cdot\rangle\!\rangle\}$, and subsequently the form of the feature-map recursion (5a) depends on this operation as specified by the operators $\square^{[j]}$ (weight map operator), $\boxdot^{[j]}$ (data map/metric operator) and $\circledcirc^{[j]}$ (norm bound operator) defined. For non-Lipschitz neural activations we set $L_a^{[\widetilde{j}:j]}=B^{[\widetilde{j}:j]}/\phi^{[\widetilde{j}]}$ in the bounded case and $L_a^{[\widetilde{j}:j]}=1$ if assumption 4 is not satisfied.

3 Neural Networks in Reproducing Kernel Banach Space

As noted in our introduction, a recurring theme in the machine learning (most famously kernel methods) is the use of bilinear (dual) representations to cleanly separate data and model parameters, ie:

$$\mathbf{f}\left(\boldsymbol{x};\Theta\right)=\left\langle \boldsymbol{\Psi}\left(\Theta\right),\boldsymbol{\phi}\left(\boldsymbol{x}\right)\right]$$

Here the set of network parameters Θ, and the data $\boldsymbol{x}$, are mapped entirely independently into distinct feature spaces by, respectively, $\boldsymbol{\Psi}:\mathbb{W}\to\mathcal{W}$ (weights and biases) and $\boldsymbol{\phi}:\mathbb{X}\to\mathcal{X}$ (data). The bilinear product $\langle\cdot,\cdot]:\mathcal{W}\times\mathcal{X}\to\mathbb{R}^m$ generalizes the inner product of eg SVMs (Cortes & Vapnik, 1995; Burges, 1998; Cristianini & Shawe-Taylor, 2005; Steinwart & Christman, 2008) without losing the very useful property of bilinearity that makes this formalism so convenient to work with. Apart from the potential for constructing a representor theory (kernelization), if the bilinear product is continuous (ie. if $\exists C,C'\in\mathbb{R}_+$ so that $\langle\boldsymbol{\Psi},\boldsymbol{\phi}]\leq C\|\boldsymbol{\phi}\|\forall\boldsymbol{\Psi}$ or $\langle\boldsymbol{\Psi},\boldsymbol{\phi}]\leq C'\|\boldsymbol{\Psi}\|\forall\boldsymbol{\phi}$) then the existence of such a model significantly simplifies the development of Rademacher complexity bounds. A model of this type was developed in (Shilton et al., 2023) using a recursive Taylor series expansion of the neural activations - in brief, noting that $\langle\mathbf{a},\mathbf{b}]_{\mathbf{g}}^{n}=\langle\mathbf{a}^{\otimes n},\mathbf{b}^{\otimes n}]_{\mathbf{g}^{\otimes n}}$, if the input to a neuron can be represented bilinearly then so too could the output, which recursion defines the model. Unfortunately this approach only works for continuous neural activations, and even then only within the RoC of the Taylor expansion, rendering it inapplicable for common activations such as ReLU. Alternatively, in this paper we propose using a Hermite polynomial expansion, which has two benefits, precisely (1) the Hermite polynomial expansion exist for all finite-energy activations and is convergent everywhere (applicability), and (2) as the Hermite polynomials are constructed from monomials we can also use $\langle\mathbf{a},\mathbf{b}]_{\mathbf{g}}^{n}=\langle\mathbf{a}^{\otimes n},\mathbf{b}^{\otimes n}]_{\mathbf{g}^{\otimes n}}$ to construct our model (practicality).

We begin by constructing our dual representation:

Theorem 1. *Let $\mathbf{f}:\mathbb{X}\times\mathbb{W}\to\mathbb{R}^m$ be a neural network (1) satisfying assumptions 1- 3. Assume nominal bounds $\|\mathbf{W}^{[\widetilde{\jmath}:\jmath]}\|_2\leq\omega^{[\widetilde{\jmath}:\jmath]}<\infty$ and $\|\mathbf{b}^{[\jmath]}\|_2\leq\beta^{[\jmath]}<\infty\;\forall j\in\mathbb{Z}_E$, $\widetilde{\jmath}\in\mathbb{A}^{[\jmath]}$. Let $\eta\in\mathbb{R}_+$. Defining feature maps $\boldsymbol{\Psi}:\mathbb{W}\to\mathcal{W}\subset\mathbb{R}^{\infty\times m}$ (weights and biases) and $\boldsymbol{\phi}:\mathbb{X}\to\mathcal{X}\subset\mathbb{R}^{\infty}$ (data) and metric $\mathbf{g}\in\mathbb{R}^{\infty}$ as per (5) (Figure 4), the network may be written in bilinear form:*

$$\mathbf{f}\left(\boldsymbol{x};\Theta\right)=\left\langle \boldsymbol{\Psi}\left(\Theta\right),\boldsymbol{\phi}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}} \tag{6}$$

where $\|\boldsymbol{\Psi}(\Theta)\|_F\leq\psi_\eta<\infty\;\forall\Theta\in\mathbb{W}$, $\|\boldsymbol{\phi}(\boldsymbol{x})\|_2\leq\phi_\eta<\infty\;\forall\boldsymbol{x}\in\mathbb{X}$, (the constants ψ_η,ϕ_η are provided in Appendix C.1), where $\lim_{\eta\to0^+}\psi_\eta=\psi$ and $\lim_{\eta\to0^+}\phi_\eta=\phi$; and we note that $\lim_{\eta\to0^+}\|\boldsymbol{\phi}(\boldsymbol{x})\|_2=\phi\;\forall\boldsymbol{x}\in\partial\mathbb{X}$ (ie. if $\|\boldsymbol{x}\|_2=1$), and $\lim_{\eta\to0^+}\|\boldsymbol{\phi}(\boldsymbol{x})\|_2>0\;\forall\boldsymbol{x}\neq\mathbf{0}$.

A full inductive proof can be found in Appendix C.1. To summarize, picking a layer $\jmath\in\mathbb{Z}_D$, we assume all nodes $\widetilde{\jmath}\in\mathbb{L}^{[\jmath-1]}$ in the preceding layer may be written $\boldsymbol{x}^{[\widetilde{\jmath}]}=\left\langle\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta),\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\widetilde{\jmath}]}}$, which is trivial for the base case $\jmath=0$. Then, using $(\mathbf{A}^{\mathrm{T}}\mathbf{b})^{\odot p}=(\mathbf{A}^{\otimes^{\updownarrow}p})^{\mathrm{T}}(\mathbf{b}^{\otimes p})$ in combination with the Hermite (number) expansion of the neural activation function, we write the incoming edge activations $\boldsymbol{x}^{[\widetilde{\jmath}:\jmath]}$ as bilinear products $\boldsymbol{x}^{[\widetilde{\jmath}:\jmath]}=\left\langle\boldsymbol{\Psi}^{[\widetilde{\jmath}:\jmath]}(\Theta),\boldsymbol{\phi}^{[\widetilde{\jmath}:\jmath]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\widetilde{\jmath}:\jmath]}}$ (see Appendix for full definitions). This, combined with the observation that $\bigcirc_{\widetilde{\jmath}\in\mathbb{A}^{[\jmath]}}^{[\jmath]}\mathbf{W}^{[\widetilde{\jmath}:\jmath]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:\jmath]}=(\square_{\widetilde{\jmath}\in\mathbb{A}^{[\jmath]}}^{[\jmath]}\mathbf{W}^{[\widetilde{\jmath}:\jmath]})^{\mathrm{T}}(\boxdot_{\widetilde{\jmath}\in\mathbb{A}^{[\jmath]}}^{[\jmath]}\mathbf{x}^{[\widetilde{\jmath}:\jmath]})$, suffices to show that $\boldsymbol{x}^{[\jmath]}=\left\langle\boldsymbol{\Psi}^{[\jmath]}(\Theta),\boldsymbol{\phi}^{[\jmath]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\jmath]}}$ as given, and the result follows by induction.

As alluded to in section 2, we can readily incorporate non-trivial nodes into this framework. In the recursive construction of the feature maps, (5a) is effectively a recipe for converting the bilinear expansion of the inputs to that node to a bilinear expansion of the node's output. As stated, (5a) is for a *trivial* node of the type shown in Figure 1, but alternatively we could wrap an entire sub-network or block *inside* this node (eg. an attention block - Figure 2b) and replace (5a) with the overall recipe for converting bilinear expansions of its input to a bilinear expansion of its output. Thus we may reasonably speak of an "attention node" in a Transformer network without needless clutter. For example Figure 2d includes a table detailing calculations for ϕ for attention, residual and LayerNorm blocks (nodes) (derivations for these can be found in Appendix D.1, D.2, D.3 and D.4).

Unfortunately the dual representation (6) is insufficient for Rademacher complexity analysis without assumption 4, which requires that the neural activations be Lipschitz or bounded (and in the latter case that $\|\boldsymbol{x}\|_2=1$). This assumption is central to casting the dual model (6) into RKBS, precisely:

Definition 1 (Reproducing kernel Banach space (RKBS))**.** A RKBS on $\mathbb{X}$ is a Banach space $\mathcal{B}$ of functions $f:\mathbb{X}\to\mathbb{Y}$, where $\mathbb{Y}$ is normed, for which the point evaluation functionals $\boldsymbol{\delta}_{\boldsymbol{x}}(f)=f(\boldsymbol{x})$ on $\mathcal{B}$ are continuous (i.e. $\forall\boldsymbol{x}\in\mathbb{X}\;\exists C_{\boldsymbol{x}}\in\mathbb{R}_+$ such that $\|\boldsymbol{\delta}_{\boldsymbol{x}}(f)\|\leq C_{\boldsymbol{x}}\|f\|_{\mathcal{B}}\;\forall f\in\mathcal{B}$).

This is somewhat generic, so following (Lin et al., 2022) we focus on the special case:

$$\mathcal{B} = \left\{ f(\cdot;\Theta) = \langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\cdot)]_{\mathcal{W}\times\mathcal{X}} \middle| \, \Theta \in \mathbb{W} \right\} \tag{7}$$

where $\boldsymbol{\phi} : \mathbb{X} \to \mathcal{X}$ is a data feature map, $\boldsymbol{\Psi} : \mathbb{W} \to \mathcal{W}$ is a weight feature map, $\mathcal{X}$ and $\mathcal{W}$ are Banach spaces, and $\langle \cdot, \cdot]_{\mathcal{W}\times\mathcal{X}} : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^m$ is a continuous bilinear form. Given this prequel we have:

Corollary 2. *The set $\mathcal{F}$ of networks (2) satisfying assumptions 1-4 with* Lipschitz *neural activations and weights and biases bounded as per Theorem 1 is an RKBS with $\|\mathbf{f}(\cdot;\Theta)\|_{\mathcal{F}} \triangleq \lim_{\eta\to 0^+} \|\boldsymbol{\Psi}(\Theta)\|_F \leq \psi < \infty$ and $\|\mathbf{f}(\boldsymbol{x};\Theta)\|_2 \leq C_{\boldsymbol{x}} \|\mathbf{f}(\cdot;\Theta)\|_{\mathcal{F}}$, where $C_{\boldsymbol{x}} = 1 \; \forall \boldsymbol{x} \in \mathbb{X}$.*

Corollary 3. *The set $\mathcal{F}^\star$ of networks (2) satisfying assumptions 1-4 with* Lipschitz *or* bounded *neural activations and with weights and biases bounded as per Theorem 1 is an RKBS with $\|\mathbf{f}(\boldsymbol{x};\cdot)\|_{\mathcal{F}^\star} \triangleq \lim_{\eta\to 0^+} \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq \phi < \infty$ and $\|\mathbf{f}(\boldsymbol{x};\Theta)\|_2 \leq C_\Theta \|\mathbf{f}(\boldsymbol{x};\cdot)\|_{\mathcal{F}^\star}$, where $C_\Theta = 1 \; \forall \Theta \in \mathbb{W}$.*

See Appendix C.3 for proofs (the structure of which minics that of the proof of Theorem 1). It follows from this that the model presented in Theorem 1 suffices to achieve our primary goal. Note that this result applies to a very wide range of networks, including feedforward ReLU networks, convolutional networks, residual networks (ResNet), and Transformer networks (see later discussion). We observe that the conditions for $\mathcal{F}$ to be an RKBS are stricter than the conditions for $\mathcal{F}^\star$ to be an RKBS, as non-Lipschitz neural activations appears incompatible with $\mathcal{F}$ being an RKBS. However as we will see that we only require $\mathcal{F}^\star$ be an RKBS to proceed with our Rademacher complexity analysis.

4 Rademacher Complexity Bounds

We now address our secondary goal, namely using our dual model to bound the Rademacher complexity of neural networks. For $h : \mathbb{R}^m \to \mathbb{R}$, the Rademacher complexity is defined as:

$$\mathcal{R}_N(h \circ \mathcal{F}) = \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\mathbf{f}\in\mathcal{F}} \tfrac{1}{N} \textstyle\sum_{i=1}^N \epsilon_i h(\mathbf{f}(\boldsymbol{x}_i)) \right]$$

for Rademacher random variables $\epsilon_i \in \{\pm 1\}$, where $\boldsymbol{x} \sim \nu$. We have:

Theorem 4. *Let $\mathcal{F}$ be the set of networks (2) satisfying assumptions 1-4 with weights and biases bounded as per Theorem 1, and let $h : \mathbb{R}^m \to \mathbb{R}$ be L-Lipschitz. Then:*

$$\mathcal{R}_N\left(h \circ \mathcal{F} : \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2 \leq \omega^{[\widetilde{j}:j]}, \left\|\mathbf{b}^{[j]}\right\|_2 \leq \beta^{[j]}\right) \leq \tfrac{H_m \phi}{\sqrt{N}} \tag{8}$$

where $H_1 = 1$ if $h = \mathrm{id}$, $H_m = \sqrt{2m}L$ otherwise, and ϕ is defined in Figure 4.

The proof follows the usual template for RKHS models (see eg. (Bartlett & Mendelson, 2002)) using our feature map; replacing the Cauchy-Schwarz inequality with $\|\mathbf{f}(\boldsymbol{x};\Theta)\|_2 \leq C_\Theta \|\boldsymbol{\phi}(\boldsymbol{x})\|_2$; taking the limit $\eta \to 0^+$; and recalling that $C_\Theta = 1$ and $\lim_{\eta\to 0^+} \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq \phi$, so $\|\mathbf{f}(\boldsymbol{x};\Theta)\|_2 \leq C_\Theta \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq \phi$. See Appendix F for full details.

Considering this Rademacher complexity bound, we recall that typically neural network weights and biases are initialized with magnitude proportional to $\frac{1}{\sqrt{H^{[j]}}}$ (LeCun initialization) or $\frac{1}{\sqrt{H^{[\widetilde{j}:j]}}}$ (He initialization), and stay close to their initial values in the wide limit, assuming a convex objective. Thus we would expect that $\|\mathbf{W}^{[\widetilde{j}:j]}\|_2$ (and hence its upper bound $\omega^{[\widetilde{j}:j]}$) should be *independent* of network width, rendering the complexity bound in Theorem 4 (effectively) width-independent. We also observe that the complexity bound does not contain any *explicitly* depth-dependent terms (nuisance terms that are often present in such bounds as discussed in (Golowich et al., 2018)); however the bound will in general grow exponentially with depth due to the multiplicative build-up of terms in ϕ from input to output, which is typical of such results (Neyshabur et al., 2015; Golowich et al., 2018; Truong, 2022). For a scalar-valued, unbiased, Lipschitz network with 1 node $j = \jmath$ per layer, (8) becomes:[3]

$$\mathcal{R}_N\left(\mathcal{F} : \left\|\mathbf{W}^{[j-1:j]}\right\|_2 \leq \omega^{[j-1:j]}\right) \leq \frac{\prod_{j=1}^D L^{[j-1:j]} \omega^{[j-1:j]}}{\sqrt{N}} \tag{9}$$

While this bound is depth-exponential in general, we can use to to derive conditions (on the weights) under which this exponentiality can be (in effect) neutralised. Motivated by this, the following result gives general, non-trivial threshold conditions for depth-independent Rademacher complexity:

[3] Conversely, we know that $\mathcal{R}_N(\mathcal{F} : \|\mathbf{W}^{[j-1:j]}\|_2 \leq \omega^{[j-1:j]}) \sim \Omega(\frac{1}{\sqrt{N}} \prod_{j=1}^D \omega^{[j-1:j]})$ (Golowich et al., 2018, Theorem 7), and hence $\mathcal{R}_N(\mathcal{F} : \left\|\mathbf{W}^{[j-1:j]}\right\|_2 \leq \omega^{[j-1:j]}) \asymp \frac{\prod_{j=1}^D \omega^{[j-1:j]}}{\sqrt{N}}$.

Node or Block Type	Depth-Independence Condition	Notes
Trivial	$\left\|\mathbf{b}^{[j]}\right\|_2^2 + \circledcirc^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_1^{[\widetilde{\jmath}:j]2} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \leq 1$	See Figure 1 and equation (10).
Residual	$\prod_{q\in\mathbb{Z}_{a_j}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]_q}\right\|_2 \leq s$	See Figure 2c. In this bound we denote the weight matrix for (internal) layer q as. $\mathbf{W}^{[\widetilde{\jmath}:j]_q}$. See Appendix D.1 for the complete derivation.
Single-Query Attention	$\lambda \left\|\mathbf{W}_V\right\|_2 \left\|\mathbf{W}_Q\right\|_2 \left\|\mathbf{W}_K\right\|_2 \leq 1$	See Figure 2a. In this bound λ is the heat parameter for the softmax. See Appendix D.3 for a complete derivation.
Single- and Multi-Head Attention	$\lambda\sqrt{d_{\text{model}}} \left\|\mathbf{W}_V\right\|_2 \left\|\mathbf{W}_Q\right\|_2 \left\|\mathbf{W}_K\right\|_2 \leq 1$	See Figure 2b. In this bound λ is the heat parameter for the softmax. Here d_{model} is the product of the number of queries and the number of heads. See Appendix D.4 for a complete derivation.

Figure 5: Conditions for Depth Independent Rademacher Complexity Bounds for Typical Nodes.

Corollary 5. *Let $\mathcal{F}$ be the set of networks (2) satisfying our assumptions with weights and biases bounded as per Theorem 1, and let $h : \mathbb{R}^m \to \mathbb{R}$ be L-Lipschitz. If:*

$$\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2 \leq \phi^{[\widetilde{\jmath}]} = 1 \forall \widetilde{\jmath} \in \mathbb{A}^{[j]} \implies \left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2 \leq \phi^{[j]} = 1 \tag{10}$$

for all nodes $j \in \mathbb{Z}_E$, then $\mathcal{R}_N(h \circ \mathcal{F}) \leq \frac{H_m}{\sqrt{N}}$, independent of both width and depth.

This follows from the recursive definition of ϕ in (5) (Figure 4) as a sufficient condition to ensure that $\phi^{[j]} = 1$ given $\phi^{[\widetilde{\jmath}]} = 1$ for all nodes $j \in \mathbb{Z}_E$, $\widetilde{\jmath} \in \mathbb{A}^{[j]}$, and subsequently (recursively) $\phi = \phi^{[E]} = 1$. In practice the interpretation of this result is node specific. Conditions for various nodes (in the Lipschitz case) can be found in Table 5, where derivations may be found in the appendices noted. The general, non-Lipschitz (bounded) case is somewhat more complicated. Recall that if there are non-Lipschitz neural activations in the network we assume that $\boldsymbol{x} \in \partial\mathbb{X}$ or, equivalently, $\|\boldsymbol{x}\|_2 = 1$; and for non-Lipschitz, bounded neural activations $\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}$, we set $L_a^{[\widetilde{\jmath}:j]} = B^{[\widetilde{\jmath}:j]}/\phi^{[\widetilde{\jmath}]}$. Considering one such non-Lipschitz neural activation $\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}$, in the recursive definition of the norm-bound ϕ in (5), the corresponding term in the sum becomes $L_a^{[\widetilde{\jmath}:j]}\phi^{[\widetilde{\jmath}]} = B^{[\widetilde{\jmath}:j]}$ - so, for example, for a LayerNorm block (Figure 2d) $j \in \mathbb{Z}_E$ we see that $\phi^{[j]} = \sqrt{H^{[\widetilde{\jmath}:j]}}$ (for full derivation see Appendix D.2), and moreover if this is the only node in its layer then the Rademacher complexity bound will be independent of all layers preceeding it. However we would advise caution here; the assumption $\boldsymbol{x} \in \partial\mathbb{X}$ is quite strong and may not be realistic in general. We will discuss how this assumption may be relaxed, along with what impact this relaxation has on our Rademacher complexity bound, in section 4.1.

4.1 Generalizations and Standard Toplogies

In this section, we consider two more realistic relaxions assumption 1 - firstly expanding the bounds on $\|\boldsymbol{x}\|_2$, and secondly considering $\boldsymbol{x} \sim \mathcal{X}$ drawn from an unbounded distribution $\mathcal{X}$ such that it lies in the bounded of case 1 with high probability (whp). Using these, we conclude the paper by analysing a range of standard network topologies. Formally, we consider two generalization of assumption 1:

Strictly Bounded: $\boldsymbol{x} \in \mathbb{X}_{\rho,r} = \{\boldsymbol{x} \in \mathbb{R}^n : \rho \leq \|\boldsymbol{x}\|_2 \leq r\}$, where $0 \leq \rho \leq r \in \mathbb{R}_+$ and $\rho > 0$ if the network contains non-Lipschitz neural activations.

Distributional: $\boldsymbol{x} \sim \mathcal{X}$ for a distribution $\mathcal{X}$ for which there exists $0 \leq \rho \leq r \in \mathbb{R}_+$ ($\rho > 0$ if the network contains non-Lipschitz neural activations) such that $\boldsymbol{x} \in \mathcal{X}_{\rho,r}$ with high probability $\geq 1 - \epsilon$.

In both cases we consider a mild modification of our feature map (5), precisely:[4]

$$\begin{aligned}
&\boldsymbol{\Psi}^{[0]}(\Theta) = r\mathbf{I}_n, \boldsymbol{\phi}^{[0]}(\boldsymbol{x}) = \boldsymbol{x}, \mathbf{g}^{[0]} = \tfrac{1}{r}\mathbf{1}_n, \psi^{[0]} = r, \phi_\downarrow^{[0]} = \rho, \phi^{[0]} = r, \qquad \phi_\downarrow = \phi_\downarrow^{[E]} \\
&\phi_\downarrow^{[j]2} = \beta^{[j]2} + \circledcirc^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \begin{cases} L_{\phi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \phi_\downarrow^{[\widetilde{\jmath}]2} & \text{if } \boldsymbol{\tau}^{[\widetilde{\jmath}:j]} \text{ is Lipschitz} \\ B^{[\widetilde{\jmath}:j]2} & \text{otherwise} \end{cases} \qquad \forall j \in \mathbb{Z}_E,
\end{aligned} \tag{11}$$

and moreover for non-Lipschitz, bounded neural activations $\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}$, we set $L_a^{[\widetilde{\jmath}:j]} = B^{[\widetilde{\jmath}:j]}/\phi_\downarrow^{[\widetilde{\jmath}]}$. For a full discussion of this generalization see Appendix C. Observe that, in the limit $\eta \to 0^+$:

$$\begin{aligned}
&\phi^{[j]2} = \beta^{[j]2} + \circledcirc^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \begin{cases} L_{\phi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \phi^{[\widetilde{\jmath}]2} & \text{if } \boldsymbol{\tau}^{[\widetilde{\jmath}:j]} \text{ is Lipschitz} \\ B^{[\widetilde{\jmath}:j]2} \frac{\phi^{[\widetilde{\jmath}]2}}{\phi_\downarrow^{[\widetilde{\jmath}]2}} & \text{otherwise} \end{cases} \qquad \forall j \in \mathbb{Z}_E \\
&\phi_\downarrow^{[j]} \leq \lim_{\eta\to 0^+} \left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2 \leq \phi^{[j]} \quad \forall \boldsymbol{x} \in \mathbb{X}_{\rho,r}
\end{aligned} \tag{12}$$

[4] In both the cases $\rho = 0$, $r = 1$ (the fully Lipschitz variant of assumption 1) and $\rho = r = 1$ (the non-Lipschitz variant of assumption 1) this reduces to the standard feature map (5).

The Rademacher complexity bound (Theorem 8) takes the same form as usual. The exact impact of letting $r \neq 1$ is dependent on the network topology. For a simple, layerwise, fully Lipschitz neural network with 1 trivial node $j = \jmath$ per layer, as demonstrated in Appendix E.1:[5]

$$\mathcal{R}_N\left(h \circ \mathcal{F} : \left\|\mathbf{W}^{[\jmath-1:\jmath]}\right\|_2 \leq \omega^{[\jmath-1:\jmath]}\right) \leq r\frac{H_m \prod_{\jmath=1}^{D} L^{[\jmath-1:\jmath]}\omega^{[\jmath-1:\jmath]}}{\sqrt{N}}$$

This bound is exponential in depth, as discussed previously. As a mild generalization of this scenario, if we allow non-Lipschitz neural activations (for example LayerNorm blocks) in this simple network, with the last such at layer $\jmath = D_\downarrow$, then, using (12) and noting that $\phi^{[\jmath]}/\phi_\downarrow^{[\jmath]} = \frac{r}{\rho}\forall \jmath \in \mathbb{Z}_D \cup \{0\}$:

$$\mathcal{R}_N\left(h \circ \mathcal{F} : \left\|\mathbf{W}^{[\jmath-1:\jmath]}\right\|_2 \leq \omega^{[\jmath-1:\jmath]}\right) \leq \frac{r}{\rho}\frac{H_m B^{[D_\downarrow-1:D_\downarrow]}\omega^{[D_\downarrow-1:D_\downarrow]}\prod_{\jmath=D_\downarrow+1}^{D} L^{[\jmath-1:\jmath]}\omega^{[\jmath-1:\jmath]}}{\sqrt{N}}$$

where we note that this bound is exponential in the depth to the non-Lipschitz node $D - D_\downarrow$ and proportional to $\frac{r}{\rho}$. The independence from the weights of layers preceeding $D_\downarrow$ is noteworthy, but if we consider as an example a ReLU network terminated by a LayerNorm and observe that the scale of these weights is entirely arbitrary, it perhaps not surprising. The $\frac{1}{\rho}$ term reflects the need to assume that, in the worst-case, small inputs will be "amplified" (e.g. by LayerNorm) to the largest possible output.

The transformer can be similarly analysed. The catch in this case is that the attention block is multiplicative. In particular (see Appendices D.3, D.4 for details), for an attention block:

$$\frac{\phi_{\text{out}}}{\phi_{\text{out}\downarrow}} = \frac{\phi_{\text{out},Q}}{\phi_{\text{out},Q\downarrow}}\frac{\phi_{\text{out},K}}{\phi_{\text{out},K\downarrow}}\frac{\phi_{\text{in},V}}{\phi_{\text{in},V\downarrow}}$$

so, unlike the simpler case considered above, each attention block will cause polynomial growth in the ratio $\frac{\phi}{\phi_\downarrow}$. Subsequently, as shown in Appendix E.3, the overall bound (due to the final LayerNorm) is:

$$\mathcal{R}_N\left(h \circ \mathcal{F} : \|\mathbf{W}_{\text{out}}\|_2 \leq \omega\right) \leq \left(\frac{\rho}{r}\right)^{3^{3M-1}} \frac{H_m\sqrt{d_{\text{model}}}\omega}{\sqrt{N}}$$

where $\mathbf{W}_{\text{out}}$ are the weights for the linear output layer of the transformer.[6] If $\rho = r$ (that is, $\boldsymbol{x} \in \partial\mathbb{X}$ as in assumption 1) this collapses to $\frac{H_m\sqrt{d_{\text{model}}}\omega}{\sqrt{N}}$, but in general, despite being independent of the weights in all but the output layer of the network, this bound grows doubly-exponentially in depth, dependent on the ratio $\frac{r}{\rho}$ of smallest/largest inputs.

Finally, bounds for the distributional case follow the strictly bounded case, but only whp $\geq 1-\epsilon$. For example, in Appendix C.4 we consider $\boldsymbol{x} \sim \mathcal{X} = \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, showing that $\rho \leq \|\boldsymbol{x}\|_2 \leq r$, where:

$$\rho = 0,\ r = \sqrt{2n\ln\left(\tfrac{2}{\epsilon}\right)}\sigma \qquad \text{or} \qquad \frac{r}{\rho} = \frac{\sqrt{n\ln\left(\frac{4}{\epsilon}\right)}}{\left(\Gamma\left(\frac{n}{2}+1\right)\frac{\epsilon}{2}\right)^{\frac{1}{n}}}$$

whp $\geq 1-\epsilon$ which apply, respectively, for the *purely Lipchitz* and *bounded* cases. In particular, the latter result allows one to explore Rademacher complexity bounds in the general case without giving ρ or r (the bounds on $\|\boldsymbol{x}\|_2$, where ρ in particular may be difficult to quantify intuitively) a-priori.

5 Conclusions

In this paper we have constructed a dual model of a very general set of feedforward neural networks that re-expresses them as a continuous bilinear product between a weight/bias feature map and a data feature map - that is, a reproducing kernel Banach space (RKBS) model. This model is *exact*, with no approximation or assumptions beyond bounded (norm) inputs, bounded (spectral norm) weights and biases, and finite-energy neural activations, and incorporates networks ranging from simple layerwise models (ReLU etc) to ResNet and Transformers. Subsequently, we have applied this model to the analysis of the Rademacher complexity analysis of neural networks, giving a simple recursive bound for the Rademacher complexity of all models neural network topologies covered by our model. This bound is exact (non-asymptotic) and does not include depth- or width- dependent nuisance factors. Moreover it is width-independent and, while exponential in depth (due to the multiplicative build-up of terms through the layers of the networks), enables us to derive straightforward (spectral) threshold conditions under which depth-dependence may be removed entirely.

[5] This also applies to ResNet, where for residual blocks $\jmath$ with d internal layers we let $\omega^{[\jmath-1:\jmath]2} = (\omega^{[\widetilde{\jmath}-1:\widetilde{\jmath}]_d 2}\ldots\omega^{[\widetilde{\jmath}-1:\widetilde{\jmath}]_2 2}\omega^{[\widetilde{\jmath}-1:\widetilde{\jmath}]_1 2} + 1 - s^2$ as also described in Appendix E.1.

[6] M here is the size of the encoder/decoder stacks. We use M here rather than N as used in (Vaswani et al., 2017) to avoid a notational ambiguity within our paper.

References

Abramowitz, M., Stegun, I. A., and McQuarrie, D. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1972.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, Jan–Jun 1950.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8139–8148, 2019b.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

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

Bartolucci, F., De Vito, E., Rosasco, L., and Vigogna, S. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, January 2023.

Boyd, J. P. The rate of convergence of hermite function series. *Mathematics of Computation*, 35 (152):1309–1316, October 1980.

Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in neural information processing systems*, volume 32, 2019.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In Y., B., D., S., D., L. J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 342–350. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf`.

Cortes, C. and Vapnik, V. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.

Courant, R. and Hilbert, D. *Methods of Mathematical Physics*. John Wiley and sons, New York, 1937.

Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2005.

Daniely, A. SGD learns the conjugate kernel class of the network. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2422–2430. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6836-sgd-learns-the-conjugate-kernel-class-of-the-network.pdf`.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2253–2261. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6427-toward-deeper-understanding-of-neural-networks-the-power-of-initialization-and-a-dual-view-on-expressivity.pdf`.

Der, R. and Lee, D. Large-margin classification in banach spaces. In *Proceedings of the JMLR Workshop and Conference 2: AISTATS2007*, pp. 91–98, 2007.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *Conference on Learning Representations*, 2019b.

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *COLT*, 2018.

Gradshteyn, I. S. and Ryzhik, I. M. *Table of Integrals, Series, and Products*. Academic Press, London, 2000.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hille, E. Contributions to the theory of Hermitian series. II. The representation problem. *Trans. Amer. Math. Soc.*, 47:80–94, 1940.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Lin, R., Zhang, H., and Zhang, J. On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 2022.

Maurer, A. A vector-contraction inequality for rademacher complexities. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 3–17, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Morse, P. M. and Feshbach, H. *Methods of Theoretical Physics*. McGraw-Hill, 1953.

Neal, R. M. *Priors for infinite networks*, pp. 29–53. Springer, 1996.

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

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Proceedings of Conference on Learning Theory*, pp. 1376–1401, 2015.

Olver, F. W., Lozier, D. W., Boisvert, R. F., and Clark, C. W. *NIST Handbook of Mathematical Functions*. Cambridge University Press, USA, 1st edition, 2010. ISBN 0521140633.

Parhi, R. and Nowak, R. D. Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.*, 22(43):1–40, 2021.

Rahman, S. Wiener-hermite polynomial expansion for multivariate gaussian probability measures. *Journal of Mathematical Analysis and Applications*, 454(1):303–334, 2017.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Sanders, K. Neural networks as functions parameterized by measures: Representer theorems and approximation benefits. Master's thesis, Eindhoven University of Technology, 2020.

Shilton, A., Gupta, S., Rana, S., and Venkatesh, S. Gradient descent in neural networks as sequential learning in reproducing kernel banach space. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31435–31488. PMLR, 23–29 Jul 2023.

Song, G., Zhang, H., and Hickernell, F. J. Reproducing kernel banach spaces with the ℓ^1 norm. *Applied and Computational Harmonic Analysis*, 34(1):96–116, Jan 2013.

Spek, L., Heeringa, T. J., and Brune, C. Duality for neural networks through reproducing kernel banach spaces. *arXiv preprint arXiv:2211.05020*, 2022.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Learning in hilbert vs. banach spaces: A measure embedding viewpoint. In *Advances in Neural Information Processing Systems*, pp. 1773–1781, 2011.

Steinwart, I. and Christman, A. *Support Vector Machines*. Springer, 2008.

Truong, L. V. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284*, 2022.

Unser, M. A representer theorem for deep neural networks. *J. Mach. Learn. Res.*, 20(110):1–30, 2019.

Unser, M. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics*, 21(4):941–960, 2021.

Valle-Pérez, G. and Louis, A. A. Generalization bounds for deep learning. *arXiv preprint arXiv:2012.04115*, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Xu, Y. and Ye, Q. Generalized mercer kernels and reproducing kernel banach spaces. *arXiv preprint arXiv:1412.8663*, 2014.

Zhang, H. and Zhang, J. Regularized learning in banach spaces as an optimization problem: representer theorems. *Journal of Global Optimization*, 54(2):235–250, Oct 2012.

Zhang, H., Xu, Y., and Zhang, J. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In *Advances in neural information processing systems*, volume 32, 2019.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

A Properties of Hermite polynomials

A.1 Univariate Case

The (probabilist's) Hermite polynomials are given by (Abramowitz et al., 1972; Morse & Feshbach, 1953; Olver et al., 2010; Courant & Hilbert, 1937):

$$He_k(\zeta) = (-1)^k e^{\frac{\zeta^2}{2}} \frac{d^k}{d\zeta^k} e^{-\frac{\zeta^2}{2}} \qquad \forall k \in \mathbb{N}$$

or, explicitly:

$$He_k(\zeta) = k! \sum_{0 \le 2p \le k} \frac{(-1)^p}{2^p p!(k-2p)!} \zeta^{k-2p} \qquad \forall k \in \mathbb{N} \tag{13}$$

and form an orthogonal basis of $L^2(\mathbb{R}, e^{-x^2})$. For any $f \in L^2(\mathbb{R}, e^{-x^2})$ there exist Hermite coefficients $a_0, a_1, \ldots \in \mathbb{R}$ (the Hermite transform of f) such that:

$$f(\zeta) = \sum_{k \in \mathbb{N}} a_k He_k(\zeta) \quad \forall \xi, \zeta \in \mathbb{R}$$

where:

$$a_k = \frac{1}{k!\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi + \zeta) e^{-\frac{\zeta^2}{2}} He_k(\zeta) \, d\zeta$$

and moreover the series representation converges everywhere on the real line.[7]

The Hermite numbers derive from the Hermite polynomials:[8]

$$\mathrm{He}_k \triangleq He_k(0) = \begin{cases} \frac{(-1)^{k/2} k!}{k!!} & \text{if } k \text{ even} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $k!! = k(k-2)(k-4)\ldots$ is the double-factorial. It is well known that (see eg. (Morse & Feshbach, 1953)):

$$He_k(\xi + \zeta) = \sum_{0 \le l \le k} \binom{k}{l} He_{k-l}(\xi) \zeta^l$$

and so:

$$He_k(\zeta) = \sum_{0 \le l \le k} \binom{k}{l} \mathrm{He}_{k-l} \zeta^l$$

It follows that, taking care not to change or order of summation (remember this is an alternating series, so convergence depends on the order of the summation):

$$f(\zeta) - f(0) = \sum_{k=1}^{\infty} a_k \sum_{l=1}^{k} \binom{k}{l} \mathrm{He}_{k-l} \zeta^l$$

For later reference we also note that the Hermite polynomials satisfy the well-known recursion and derivative relation for $k > 1$:

$$\begin{aligned} \zeta H_k(\zeta) &= \tfrac{1}{2} H_{k+1}(\zeta) + \tfrac{1}{2} H_k'(\zeta) \\ &= \tfrac{1}{2} H_{k+1}(\zeta) + k H_{k-1}(\zeta) \end{aligned} \tag{16}$$

[7] Hille (1940); Boyd (1980) show that this series converges on a strip $\mathbb{X}_\rho = \{z \in \mathbb{C} : -\rho < \mathrm{Im}(z) < \rho\}$ of width ρ about the real axis in the complex plane, where (note that Hille (1940); Boyd (1980) use the normalized physicist's Hermite polynomials. The additional scale factor here arises in the translation to the un-normalized probabilist's Hermite polynomials used here):

$$\rho = -\limsup_{k \to \infty} \frac{1}{\sqrt{2k+1}} \log\left(\left| \frac{a_k}{\sqrt{k!\sqrt{\pi}}} \right| \right) \tag{14}$$

[8] Typically the Hermite numbers are defined from the physicist's Hermite polynomials, but as we use the Probabilist's form as we find these more convenient for our purposes.

A.2 Multivariate Case

The multivariate Hermite polynomials $He_{\mathtt{k}} : \mathbb{R}^n \to \mathbb{R}$, $\mathtt{k} \in \mathbb{N}^n$, are the functions (Rahman, 2017):

$$He_{\mathtt{k}}(\boldsymbol{\zeta}) = (-1)^{|\mathtt{k}|} \exp\left(\tfrac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\zeta}\right) \tfrac{\partial^{\mathtt{k}}}{\partial \boldsymbol{\zeta}^{\mathtt{k}}} \exp\left(-\tfrac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\zeta}\right) = \textstyle\prod_i He_{k_i}(\zeta_i)$$

where we use multi-index notation $|\mathtt{k}| = \sum_i k_i$, $\mathbf{a}^{\mathtt{k}} = \prod_i a_i^{k_i}$, $\mathtt{k}! = \prod_i k_i!$, $\mathtt{k}!! = \prod_i k_i!!$, and $\frac{\partial^{\mathtt{k}}}{\partial \boldsymbol{\zeta}^{\mathtt{k}}} = \prod_i \frac{\partial^{k_i}}{\partial \zeta_i^{k_i}}$. For any $f \in L^2(\mathbb{R}^n, e^{-\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\zeta}})$ there exists coefficients $a_{\mathtt{k}} \in \mathbb{R} : \mathtt{k} \succeq_n i$ (the Hermite transform of f), where $\mathtt{k} \succeq_n i$ means $\mathtt{k} \in \{\mathtt{k} \in \mathbb{N}^n : |\mathtt{k}| \geq i\}$, such that:

$$f(\boldsymbol{\zeta}) = \sum_{\mathtt{k} \succeq_n 0} a_{\mathtt{k}} He_{\mathtt{k}}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^n$$

where:

$$\begin{aligned} a_{\boldsymbol{k}} &= \tfrac{1}{\mathtt{k}!(2\pi)^{\frac{n}{2}}} \textstyle\int_{-\infty}^{\infty} f(\boldsymbol{\zeta})\, e^{-\frac{\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{\zeta}}{2}} He_{\mathtt{k}}(\boldsymbol{\zeta})\, d\boldsymbol{\zeta} \\ &= \tfrac{1}{k_1!\sqrt{2\pi}} \textstyle\int_{\zeta_1 \in \mathbb{R}} e^{-\frac{\zeta_1^2}{2}} He_{k_1}(\zeta_1) \tfrac{1}{k_2!\sqrt{2\pi}} \int_{\zeta_2 \in \mathbb{R}} e^{-\frac{\zeta_2^2}{2}} He_{k_2}(\zeta_2) \ldots f(\boldsymbol{\zeta}) \ldots d\zeta_2 d\zeta_1 \end{aligned}$$

and the series representation converges everywhere on $\mathbb{R}^n$.

As in the univariate case, the multivariate Hermite numbers are defined as:

$$\mathrm{He}_{\mathtt{k}} \triangleq He_{\mathtt{k}}(\mathbf{0}) = \textstyle\prod_i \mathrm{He}_{k_i} = \begin{cases} \frac{(-1)^{|\mathtt{k}|/2}\mathtt{k}!}{\mathtt{k}!!} & \text{if } k_0, k_1, \ldots \text{ are all even} \\ 0 & \text{otherwise} \end{cases}$$

where in the final step we have used (15). Subsequently:

$$f(\boldsymbol{\zeta}) - f(\mathbf{0}) = \sum_{\mathtt{k} \succ_n 0} a_{\mathtt{k}} \sum_{0 \prec_n \mathtt{l} \leq \mathtt{k}} \tbinom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \boldsymbol{\zeta}^{\mathtt{l}}$$

where $\mathtt{k} \succ_n i$ means $\mathtt{k} \in \{\mathtt{k} \in \mathbb{N}^n : |\mathtt{k}| > i\}$.

Finally, if we consider a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ then it is not hard to see that scalar-valued expansion can be extended to:

$$\mathbf{f}(\boldsymbol{\zeta}) - \mathbf{f}(\mathbf{0}) = \sum_{\mathtt{k} \succ_n 0} \mathbf{a}_{\mathtt{k}} \sum_{0 \prec_n \mathtt{l} \leq \mathtt{k}} \tbinom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \boldsymbol{\zeta}^{\mathtt{l}} \tag{17}$$

where $a_{\mathtt{k},i}$ are the Hermite coefficients of f_i. We note that if $n = m$ and $\mathbf{f}(\boldsymbol{\zeta}) = [g(\zeta_i)]_i$ acts elementwise (for example a neural activation that acts elementwise) then:

$$a_{\mathtt{k},i} = \delta_{|\mathtt{k}|,k_i} b_{|\mathtt{k}|} \tag{18}$$

where $b_0, b_1, \ldots$ are the (univariate) Hermite coefficients of $g : \mathbb{R} \to \mathbb{R}$.

B ReLU Activation Function Analysis

In this section we derive the Hermite-polynomial expansion of the ReLU activation function:

$$\tau^{[\mathrm{ReLU}]}(\zeta) = [\zeta]_+$$

We find it convenient to work in terms of the physicists Hermite polynomials H_k to suit (Gradshteyn & Ryzhik, 2000). So:

$$\begin{aligned} b_k^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi k!}} \textstyle\int_0^{\infty} \zeta e^{-\frac{\zeta^2}{2}} He_k(\zeta)\, d\zeta \\ &= \tfrac{1}{\sqrt{2\pi k!}} \textstyle\int_0^{\infty} \zeta e^{-\frac{\zeta^2}{2}} \frac{1}{\sqrt{2^k}} H_k\left(\frac{\zeta}{\sqrt{2}}\right) d\zeta \\ &= \sqrt{\tfrac{2}{\pi}} \tfrac{1}{k!} \textstyle\int_0^{\infty} \frac{\zeta}{\sqrt{2}} e^{-\left(\frac{\zeta}{\sqrt{2}}\right)^2} \frac{1}{\sqrt{2^k}} H_k\left(\frac{\zeta}{\sqrt{2}}\right) d\frac{\zeta}{\sqrt{2}} \\ &= \sqrt{\tfrac{2}{\pi}} \tfrac{1}{\sqrt{2^k} k!} \textstyle\int_0^{\infty} \zeta e^{-\zeta^2} H_k(\zeta)\, d\zeta \end{aligned}$$

and hence, using (16) and (Gradshteyn & Ryzhik, 2000, (7.373)):

$$b_k^{[\mathrm{ReLU}]} = \tfrac{k+1}{\sqrt{\pi}} \tfrac{1}{\sqrt{2}^{k+1}(k+1)!} \textstyle\int_0^{\infty} e^{-\zeta^2} H_{k+1}(\zeta)\, d\zeta + \tfrac{1}{\sqrt{\pi}} \tfrac{1}{\sqrt{2}^{k-1}(k-1)!} \int_0^{\infty} e^{-\zeta^2} H_{k-1}(\zeta)\, d\zeta$$

and using (Gradshteyn & Ryzhik, 2000, (7.373)) again:

$$\begin{aligned} b_k^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{\pi}} \tfrac{1}{\sqrt{2}^{k+1}(k+1)!}(k+1)\left(e^0 H_k\left(0\right) - e^{-\frac{\infty^2}{2}} H_k\left(\tfrac{\infty}{\sqrt{2}}\right)\right) + \dots \\ &\qquad \dots \tfrac{1}{\sqrt{\pi}} \tfrac{1}{\sqrt{2}^{k-1}(k-1)!}\left(e^0 H_{k-2}\left(0\right) - e^{-\frac{\infty^2}{2}} H_{k-2}\left(\tfrac{\infty}{\sqrt{2}}\right)\right) \\ &= \tfrac{1}{\sqrt{2\pi}}\left(\tfrac{k+1}{\sqrt{2}^k(k+1)!} H_k\left(0\right) + \tfrac{1}{\sqrt{2}^{k-2}(k-1)!} H_{k-2}\left(0\right)\right) \end{aligned}$$

If $k = 2p$ and $p > 0$ then, noting that $H_k(0) = \sqrt{2}^k \mathrm{He}_k$:

$$\begin{aligned} b_{2p}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi}}\left(\tfrac{1}{\sqrt{2}^{2p}(2p+1)!}(2p+1)H_{2p}\left(0\right) + \tfrac{1}{\sqrt{2}^{2p-2}(2p-1)!} H_{2p-2}\left(0\right)\right) \\ &= \tfrac{1}{\sqrt{2\pi}}\left(\tfrac{1}{\sqrt{2}^{2p}(2p+1)!}(2p+1)H_{2p}\left(0\right) + \tfrac{1}{\sqrt{2}^{2p-2}(2p-1)!} H_{2p-2}\left(0\right)\right) \\ &= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!}\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!} H_{2p}\left(0\right) + \tfrac{(-1)^{p+1}2p!}{(2(p-1))!} H_{2p-2}\left(0\right)\right) \\ = \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!} e^{-\frac{\xi^2}{2}}&\left(\tfrac{(-1)^{p+1}(2p-1)p!}{(2p)!} 2^p \tfrac{(-1)^p(2p)!}{2^p p!} + \tfrac{(-1)^{p+1}2p!}{(2(p-1))!} 2^{p-1} \tfrac{(-1)^{p+1}(2p-2)!}{(p-1)!2^{p-1}}\right) \\ &= \tfrac{(-1)^{p+1}}{\sqrt{2\pi}(2p-1)2^p p!} \end{aligned}$$

If $k = 2p+1$ and $p > 0$ then:

$$\begin{aligned} b_{2p+1}^{[\mathrm{ReLU}]} &= \tfrac{1}{\sqrt{2\pi}}\left(\tfrac{1}{\sqrt{2}^{2p+1}(2p+2)!}(2p+2)H_{2p+1}\left(0\right) + \tfrac{1}{\sqrt{2}^{2p-1}(2p)!} H_{2p-1}\left(0\right)\right) \\ &= \tfrac{1}{\sqrt{2\pi}}\left(-\tfrac{1}{\sqrt{2}^{2p+1}(2p+2)!}(2p+2)H_{2p+1}\left(0\right) - \tfrac{1}{\sqrt{2}^{2p-1}(2p)!} H_{2p-1}\left(0\right)\right) \\ &= \tfrac{1}{\sqrt{2\pi}}\left(-\tfrac{\sqrt{2}}{2^{p+1}(2p+1)!} H_{2p+1}\left(0\right) - \tfrac{\sqrt{2}}{2^p(2p)!} H_{2p-1}\left(0\right)\right) \\ &= 0 \end{aligned}$$

For the cases $k = 0, 1$ we use the result:

$$\textstyle\int_a^b \zeta^m e^{-\zeta^2} d\zeta = \frac{1}{2}\Gamma\left(\frac{m+1}{2}, a^2\right) - \frac{1}{2}\Gamma\left(\frac{m+1}{2}, b^2\right)$$

and so:

$$\textstyle\int_a^\infty \zeta^m e^{-\zeta^2} d\zeta = \frac{1}{2}\Gamma\left(\frac{m+1}{2}, a^2\right)$$

In the case $k = 0$:

$$\begin{aligned} b_0^{[\mathrm{ReLU}]} &= \sqrt{\tfrac{2}{\pi}} \textstyle\int_0^\infty \zeta e^{-\zeta^2} d\zeta \\ &= \tfrac{1}{\sqrt{2\pi}}\Gamma\left(1, 0\right) \\ &= \tfrac{1}{\sqrt{2\pi}} \end{aligned}$$

and in the case $k = 1$:

$$\begin{aligned} b_1^{[\mathrm{ReLU}]} &= \tfrac{2}{\sqrt{\pi}} \textstyle\int_0^\infty \zeta^2 e^{-\zeta^2} d\zeta \\ &= \tfrac{1}{\sqrt{\pi}}\Gamma\left(\tfrac{3}{2}, 0\right) \\ &= \tfrac{1}{2} \end{aligned}$$

Subsequently, for the elementwise ReLU neural activaiton, using (18):

$$\mathbf{a_k} = \left[\ \delta_{|\mathbf{k}|,k_i} b_{|\mathbf{k}|}\ \right]_i \tag{19}$$

Next we derive the magnitude functions for the ReLU. Using integration by parts, we see that:

$$\begin{aligned} \tfrac{1}{\sqrt{2\pi}} \textstyle\int_c^\zeta \frac{1}{\xi^2}\left(e^{\frac{1}{2}\xi^2} - 1\right) d\xi &= \tfrac{1}{\sqrt{2\pi}} \tfrac{1}{\sqrt{2}} \textstyle\int_c^\zeta \frac{2}{\xi^2}\left(e^{\frac{1}{2}\xi^2} - 1\right) d\frac{\xi}{\sqrt{2}} \\ &= \tfrac{1}{\sqrt{2\pi}} \tfrac{1}{\sqrt{2}} \textstyle\int_{\frac{c}{\sqrt{2}}}^{\frac{\zeta}{\sqrt{2}}} \frac{1}{\xi^2}\left(e^{\xi^2} - 1\right) d\xi \\ &= -\tfrac{1}{\sqrt{2\pi}} \tfrac{1}{\zeta}\left(e^{\frac{1}{2}\zeta^2} - 1\right) + \tfrac{1}{\sqrt{2\pi}} \tfrac{1}{c}\left(e^{\frac{1}{2}c^2} - 1\right) + \tfrac{1}{\sqrt{\pi}} \textstyle\int_{\frac{c}{\sqrt{2}}}^{\frac{\zeta}{\sqrt{2}}} e^{\xi^2} d\xi \\ &= -\tfrac{1}{\sqrt{2\pi}} \tfrac{1}{\zeta}\left(e^{\frac{1}{2}\zeta^2} - 1\right) + \tfrac{1}{\sqrt{2\pi}} \tfrac{1}{c}\left(e^{\frac{1}{2}c^2} - 1\right) + \tfrac{1}{2}\tfrac{2}{\sqrt{\pi}} \textstyle\int_{\frac{c}{\sqrt{2}}}^{\frac{\zeta}{\sqrt{2}}} e^{\xi^2} d\xi \\ = -\tfrac{1}{\sqrt{2\pi}} \tfrac{1}{\zeta}\left(e^{\frac{1}{2}\zeta^2} - 1\right) + \tfrac{1}{2}&\mathrm{erfi}\left(\tfrac{\zeta}{\sqrt{2}}\right) - \tfrac{1}{2}\left(\mathrm{erfi}\left(\tfrac{c}{\sqrt{2}}\right) - \tfrac{1}{\sqrt{2\pi}}\tfrac{2}{c}\left(e^{\frac{1}{2}c^2} - 1\right)\right) \end{aligned}$$

So:

$$\begin{aligned}
\sum_{k=1}^{\infty}\left|b_k^{[\mathrm{ReLU}]}\right|\zeta^k &= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\sum_{p=1}^{\infty}\tfrac{\zeta^{2p}}{(2p-1)2^p p!}\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\zeta\sum_{p=1}^{\infty}\tfrac{\zeta^{2p-1}}{(2p-1)2^p p!}\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\zeta\int_c^{\zeta}\left(\tfrac{\partial}{\partial\xi}\sum_{p=1}^{\infty}\tfrac{\xi^{2p-1}}{(2p-1)2^p p!}\right)d\xi\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\zeta\int_c^{\zeta}\left(\sum_{p=1}^{\infty}\tfrac{\xi^{2p-2}}{2^p p!}\right)d\xi\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{2\sqrt{2\pi}}\zeta\int_c^{\zeta}\left(\sum_{p=1}^{\infty}\tfrac{1}{p!}\left(\tfrac{1}{2}\xi^2\right)^{p-1}\right)d\xi\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\zeta\int_c^{\zeta}\tfrac{1}{\xi^2}\left(\sum_{p=1}^{\infty}\tfrac{1}{p!}\left(\tfrac{1}{2}\xi^2\right)^{p}\right)d\xi\\
&= \tfrac{1}{2}\zeta+\tfrac{1}{\sqrt{2\pi}}\zeta\int_c^{\zeta}\tfrac{1}{\xi^2}\left(e^{\frac{1}{2}\xi^2}-1\right)d\xi\\
&= \tfrac{1}{2}\zeta\left(\mathrm{erfi}\left(\tfrac{\zeta}{\sqrt{2}}\right)+1-\mathrm{erfi}\left(\tfrac{c}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2\pi}}\tfrac{2}{c}\left(e^{\frac{1}{2}c^2}-1\right)\right)+\tfrac{1}{\sqrt{2\pi}}\left(1-e^{\frac{1}{2}\zeta^2}\right)
\end{aligned}$$

Select c so that the first derivative is $\frac{1}{2}\zeta$:

$$-\mathrm{erfi}\left(\tfrac{c}{\sqrt{2}}\right)+\tfrac{1}{\sqrt{2\pi}}\tfrac{2}{c}\left(e^{\frac{1}{2}c^2}-1\right)=0 \text{ if } c=0$$

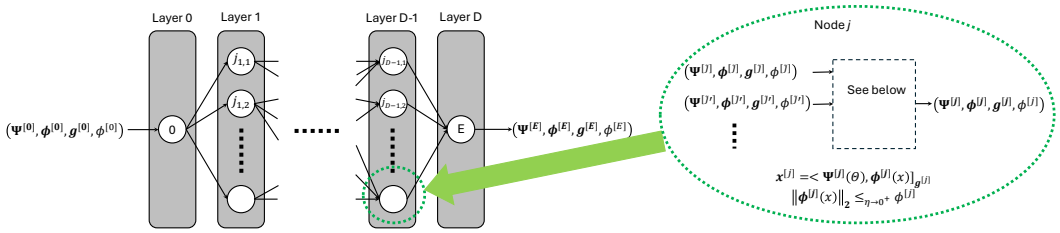
Hence:

$$\begin{aligned}
&s_{\eta}^{[\mathrm{ReLU}]}(\zeta)\triangleq\sum_{k=1}^{\infty}\left|b_k^{[\mathrm{ReLU}]}\right|(1+\eta\zeta)^k-\sum_{k=1}^{\infty}\left|b_k^{[\mathrm{ReLU}]}\right|\\
&=\tfrac{1}{2}(1+\eta\zeta)\left(\mathrm{erfi}\left(\tfrac{1+\eta\zeta}{\sqrt{2}}\right)+1\right)+\tfrac{1}{\sqrt{2\pi}}\left(1-e^{\frac{1}{2}(1+\eta\zeta)^2}\right)-\tfrac{1}{2}\left(\mathrm{erfi}\left(\tfrac{1}{\sqrt{2}}\right)+1\right)-\dots\\
&\qquad\qquad\dots\tfrac{1}{\sqrt{2\pi}}\left(1-e^{\frac{1}{2}}\right)\\
&=\tfrac{1}{2}\eta\zeta\left(\mathrm{erfi}\left(\tfrac{1+\eta\zeta}{\sqrt{2}}\right)+1\right)+\tfrac{1}{\sqrt{2\pi}}\left(e^{\frac{1}{2}}-e^{\frac{1}{2}(1+\eta\zeta)^2}\right)+\tfrac{1}{2}\left(\mathrm{erfi}\left(\tfrac{1+\eta\zeta}{\sqrt{2}}\right)-\mathrm{erfi}\left(\tfrac{1}{\sqrt{2}}\right)\right)
\end{aligned}\tag{20}$$

C Bilinear Representation - Proofs, Bounds and Generalizations

In this section we present proof of theorems, bounds and generalizations related to the bilinear representation. To avoid repeating work we consider a mild generalization of the map presented in the body of the paper, as shown in Figure 6. The key generalizations here over the main body of the paper are:

1. We let $\boldsymbol{x}\in\mathbb{X}_{\rho,r}=\{\boldsymbol{x}\in\mathbb{R}^n:\rho\leq\|\boldsymbol{x}\|_2\leq r\}$ for some $0\leq\rho\leq r\in\mathbb{R}_+$. In the main body of the paper we let $\rho=0$, $r=1$ for simplicity when all neural activations are Lipschitz, and $\rho=r=1$ otherwise. In general we require $\rho>0$ when considering a network containing non-Lipschitz neural activations.
2. We use base-case $\boldsymbol{\Psi}^{[0]}(\Theta)=r\mathbf{I}_n$, $\mathbf{g}^{[0]}=\frac{1}{r}\mathbf{1}_n$ here (recall $r=1$ in the main body).
3. We use $L_{\psi_\eta^{[\vec{\jmath}]}}^{[\vec{\jmath}:j]}$, $L_{\phi_\eta^{[\vec{\jmath}]}}^{[\vec{\jmath}:j]}$ to scale the feature map here rather than $L_{\psi^{[\vec{\jmath}]}}^{[\vec{\jmath}:j]}$ and $L_{\phi^{[\vec{\jmath}]}}^{[\vec{\jmath}:j]}$. Note, however, that (as we demonstrate) $\lim_{\eta\to0}\psi_\eta^{[\vec{\jmath}]}=\psi^{[\vec{\jmath}]}$ and $\lim_{\eta\to0}\phi_\eta^{[\vec{\jmath}]}=\phi^{[\vec{\jmath}]}$, so the definitions coincide in the limit $\eta\to0^+$, which is the case we are primarily concerned with (as it is used in our Rademacher complexity bound).
4. For non-Lipschitz, bounded neural activations (edges), we let $L^{[\vec{\jmath}:j]}=\frac{B^{[\vec{\jmath}:j]}}{\phi_{\downarrow\eta}^{[\vec{\jmath}]2}}$, where $\phi_{\downarrow\eta}^{[\vec{\jmath}]}$ is a lower bound on $\|\boldsymbol{\phi}^{[\vec{\jmath}]}(\boldsymbol{x})\|_2$ (recall that $\rho=1$ in the main body of the paper, and note that we will prove that $\phi_{\downarrow\eta}^{[\vec{\jmath}]}=\phi_\eta^{[\vec{\jmath}]}$ in this case). More generally for neural activations that are neither bounded or Lipschitz we let $L^{[\vec{\jmath}:j]}=1$. Note, however, that we cannot prove continuity of our bilinear product in this case, so the relevant parts of the proof do not apply for this.



$$\boldsymbol{\Psi}^{[0]}(\Theta) = r\mathbf{I}_n, \boldsymbol{\phi}^{[0]}(\boldsymbol{x}) = \boldsymbol{x}, \mathbf{g}^{[0]} = \tfrac{1}{r}\mathbf{1}_n$$

$$\begin{matrix}\psi_\eta^{[0]} = r \\ \phi_{\downarrow\eta}^{[0]} = \rho \\ \psi_\eta^{[0]} = r\end{matrix} \quad \underset{\eta\to 0^+}{\longrightarrow} \quad \begin{matrix}\psi^{[0]} = r \\ \phi_\downarrow^{[0]} = \rho \\ \psi^{[0]} = r\end{matrix}$$

$$\downarrow$$

$$\left.\begin{aligned}
\boldsymbol{\Psi}^{[\widetilde{j}:j]}(\Theta) &= \frac{L^{[\widetilde{j}:j]}}{\sqrt{s_\eta^{[\widetilde{j}:j]}(1)}}\left[\begin{array}{c}\eta\mathrm{diag}\left(\boldsymbol{\tau}^{[\widetilde{j}:j]}(\mathbf{0})^{\langle\odot\frac{1}{2}\rangle}\right) \\ \left[\mathrm{diag}\left(\mathbf{a}_{\mathbf{k}}^{[\widetilde{j}:j]\langle\odot\frac{1}{2}\rangle}\right)\otimes^{\updownarrow}\left[\binom{\mathbf{k}}{\mathbf{1}}^{\frac{1}{2}}\bigotimes_{\widetilde{i}\in\mathbb{N}_{H^{[\widetilde{j}]}}}\left(\sqrt{\eta}\boldsymbol{\Psi}_{\widetilde{i}\mathbf{i}}^{[\widetilde{j}]}(\Theta)\right)^{\otimes l_{\widetilde{i}}}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0}\end{array}\right] \\
\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x}) &= \frac{L^{[\widetilde{j}:j]}}{\sqrt{s_\eta^{[\widetilde{j}:j]}(1)}}\left[\begin{array}{c}\eta\left|\boldsymbol{\tau}^{[\widetilde{j}:j]}(\mathbf{0})\right|^{\odot\frac{1}{2}} \\ \left[\left|\mathbf{a}_{\mathbf{k}}^{[\widetilde{j}:j]}\right|^{\odot\frac{1}{2}}\otimes\left[\binom{\mathbf{k}}{\mathbf{1}}^{\frac{1}{2}}\left(\sqrt{\eta}\boldsymbol{\phi}^{[\widetilde{j}]}(\boldsymbol{x})\right)^{\otimes|\mathbf{l}|}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0}\end{array}\right] \\
\mathbf{g}^{[\widetilde{j}:j]} &= \frac{s_\eta^{[\widetilde{j}:j]}(1)}{L^{[\widetilde{j}:j]}_{\psi_\eta^{[\widetilde{j}]}}L^{[\widetilde{j}:j]}_{\phi_\eta^{[\widetilde{j}]}}}\left[\begin{array}{c}\frac{1}{\eta^2}\mathbf{1}_{H^{[\widetilde{j}:j]}} \\ \left[\mathbf{1}_{H^{[\widetilde{j}:j]}}\otimes\left[\mathrm{He}_{\mathbf{k}-\mathbf{1}}\left(\frac{1}{\eta}\boldsymbol{g}^{[\widetilde{j}]}\right)^{\otimes|\mathbf{l}|}\right]_{0\prec_{H^{[\widetilde{j}]}}\mathbf{1}\leq\mathbf{k}}\right]_{\mathbf{k}\succ_{H^{[\widetilde{j}]}}0}\end{array}\right]
\end{aligned}\right\} \forall \widetilde{j}\in\mathbb{A}^{[j]} \text{ (edges)} \quad \text{(a)}$$

$$\begin{matrix}\psi_\eta^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\psi_\eta^{[\widetilde{j}]}}\frac{s_\eta^{[\widetilde{j}:j]}\left(\psi_\eta^{[\widetilde{j}]2}\right)}{s_\eta^{[\widetilde{j}:j]}(1)} \\ \phi_{\downarrow\eta}^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\phi_\eta^{[\widetilde{j}]}}\frac{s_\eta^{[\widetilde{j}:j]}\left(\phi_{\downarrow\eta}^{[\widetilde{j}]2}\right)}{s_\eta^{[\widetilde{j}:j]}(1)} \\ \phi_\eta^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\phi_\eta^{[\widetilde{j}]}}\frac{s_\eta^{[\widetilde{j}:j]}\left(\phi_\eta^{[\widetilde{j}]2}\right)}{s_\eta^{[\widetilde{j}:j]}(1)}\end{matrix} \quad \underset{\eta\to 0^+}{\longrightarrow} \quad \begin{matrix}\psi^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\psi^{[\widetilde{j}]}}\psi^{[\widetilde{j}]2} \\ \phi_\downarrow^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\phi^{[\widetilde{j}]}}\phi_\downarrow^{[\widetilde{j}]2} \\ \phi^{[\widetilde{j}:j]2} = L^{[\widetilde{j}:j]2}_{\phi^{[\widetilde{j}]}}\phi^{[\widetilde{j}]2}\end{matrix}$$

$$\downarrow$$

$$\left.\begin{aligned}
\boldsymbol{\Psi}^{[j]}(\Theta) &= \left[\begin{array}{c}\mathbf{b}^{[j]\mathrm{T}} \\ \square^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}}\boldsymbol{\Psi}^{[\widetilde{j}:j]}(\Theta)\mathbf{W}^{[\widetilde{j}:j]}\end{array}\right] \\
\boldsymbol{\phi}^{[j]}(\boldsymbol{x}) &= \left[\begin{array}{c}\beta^{[j]} \\ \boxdot^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}}\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\omega^{[\widetilde{j}:j]}\end{array}\right] \\
\mathbf{g}^{[j]} &= \left[\begin{array}{c}\frac{1}{\beta^{[j]}}\gamma^{[j]} \\ \boxdot^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}}\frac{1}{\omega^{[\widetilde{j}:j]}}\mathbf{g}^{[\widetilde{j}:j]}\end{array}\right]
\end{aligned}\right\} \forall j\in\mathbb{L}^{[j]} \text{ (block)} \quad \text{(b)} \quad \text{Layers } \boldsymbol{j}\in\mathbb{Z}_D \text{ (sequentially)} \qquad (21)$$

$$\begin{matrix}\psi_\eta^{[j]2} = \beta^{[j]2} + \circledcirc^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\psi_\eta^{[\widetilde{j}:j]2} \\ \phi_{\downarrow\eta}^{[j]2} = \beta^{[j]2} + \circledcirc_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\phi_{\downarrow\eta}^{[\widetilde{j}:j]2} \\ \phi_\eta^{[j]2} = \beta^{[j]2} + \circledcirc_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\phi_\eta^{[\widetilde{j}:j]2}\end{matrix} \quad \underset{\eta\to 0^+}{\longrightarrow} \quad \begin{matrix}\psi^{[j]2} = \beta^{[j]2} + \circledcirc^{[j]}_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\psi^{[\widetilde{j}:j]2} \\ \phi_\downarrow^{[j]2} = \beta^{[j]2} + \circledcirc_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\phi_\downarrow^{[\widetilde{j}:j]2} \\ \phi^{[j]2} = \beta^{[j]2} + \circledcirc_{\widetilde{j}\in\mathbb{A}^{[j]}}\omega^{[\widetilde{j}:j]2}\phi^{[\widetilde{j}:j]2}\end{matrix}$$

$$\downarrow$$

$$\boldsymbol{\Psi}(\Theta) = \boldsymbol{\Psi}^{[E]}(\Theta), \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}^{[E]}(\boldsymbol{x}), \mathbf{g} = \mathbf{g}^{[E]}$$

$$\begin{matrix}\psi_\eta = \psi_\eta^{[E]} \\ \phi_{\downarrow\eta} = \phi_{\downarrow\eta}^{[E]} \\ \phi_\eta = \phi_\eta^{[E]}\end{matrix} \quad \underset{\eta\to 0^+}{\longrightarrow} \quad \begin{matrix}\psi = \psi^{[E]} \\ \phi_\downarrow = \phi_\downarrow^{[E]} \\ \phi = \phi^{[E]}\end{matrix}$$

Where: $\square^{[j]} = \begin{cases}\mathrm{diag} & \text{if } \bigcirc^{[j]} = \bigoplus \\ \bigoplus & \text{if } \bigcirc^{[j]} = \sum \\ \bigotimes & \text{if } \bigcirc^{[j]} = \bigotimes \\ \bigotimes^{\updownarrow} & \text{if } \bigcirc^{[j]} = \odot \\ (\bigotimes^{\updownarrow}\cdot)\mathbf{1} & \text{if } \bigcirc^{[j]} = \langle\!\langle\cdot\rangle\!\rangle\end{cases}$, $\boxdot^{[j]} = \begin{cases}\bigoplus & \text{if } \bigcirc^{[j]}\in\{\bigoplus,\sum\} \\ \bigotimes & \text{if } \bigcirc^{[j]}\in\{\bigotimes,\odot,\langle\!\langle\cdot\rangle\!\rangle\}\end{cases}$ and $\circledcirc^{[j]} = \begin{cases}\sum & \text{if } \bigcirc^{[j]}\in\{\bigoplus,\sum\} \\ \prod & \text{if } \bigcirc^{[j]}\in\{\bigotimes,\odot,\langle\!\langle\cdot\rangle\!\rangle\}\end{cases}$

Figure 6: Complete version of Figure 4 (recursive definition of bilinear representation) splitting edge/node maps, showing limits and including correct weights (the main body uses simplified weights that are correct in the limit and sets $r = 1$, $\rho = 0$ (or $\rho = 1$ if non-Lipschitz neurons are present). For non-Lipschitz, bounded neural activations $\boldsymbol{\tau}^{[\widetilde{j}:j]}$ we set $L_a^{[\widetilde{j}:j]} = B^{[\widetilde{j}:j]}/\phi_{\downarrow\eta}^{[\widetilde{j}]2}$, and for non-Lipschitz and unbounded neural activations we set $L_a^{[\widetilde{j}:j]} = 1$.

Note that for each $j \in \mathbb{Z}_E$ the feature map construction is split into two steps - a construction (21a) for the incoming edges $[\widetilde{\jmath} : j]$, which we refer to as the **edge case**; and a construction (21b) for the (core of the) node itself, which we refer to as the **node case**. This split will simplify our proofs and improve clarity by separating the key steps therein. As in the main body of the paper the overall representation is:

$$\mathbf{f}(\boldsymbol{x}; \Theta) = \left\langle \boldsymbol{\Psi}(\Theta), \boldsymbol{\phi}(\boldsymbol{x}) \right]_{\mathbf{g}} \tag{22}$$

We will also show that:

$$\begin{aligned} \boldsymbol{x}^{[\widetilde{\jmath}:j]} &= \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}} \quad \forall j \in \mathbb{Z}_E, \widetilde{\jmath} \in \mathbb{A}^{[j]} \\ \boldsymbol{x}^{[j]} &= \left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[j]}} \quad \forall j \in \mathbb{Z}_E \cup \{0\} \end{aligned} \tag{23}$$

where the following bounds hold:

$$\begin{aligned} \left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F &\leq \psi_\eta^{[\widetilde{\jmath}:j]} && \forall j \in \mathbb{Z}_E, \widetilde{\jmath} \in \mathbb{A}^{[j]}, \Theta \in \mathbb{W} \\ \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2 &\in \left[\phi_{\eta\downarrow}^{[\widetilde{\jmath}:j]}, \phi_\eta^{[\widetilde{\jmath}:j]}\right] && \forall j \in \mathbb{Z}_E, \widetilde{\jmath} \in \mathbb{A}^{[j]}, \boldsymbol{x} \in \mathbb{W} \\ \left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F &\leq \psi_\eta^{[j]} && \forall j \in \mathbb{Z}_E \cup \{0\}, \Theta \in \mathbb{W} \\ \left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2 &\in \left[\phi_{\eta\downarrow}^{[j]}, \phi_\eta^{[j]}\right] && \forall j \in \mathbb{Z}_E \cup \{0\}, \boldsymbol{x} \in \mathbb{X} \end{aligned} \tag{24}$$

noting that $\phi_{\eta\downarrow}^{[\widetilde{\jmath}:j]}, \phi_{\eta\downarrow}^{[j]} > 0$ if $\rho > 0$ and:

$$\begin{aligned} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2 &= \phi_{\eta\downarrow}^{[\widetilde{\jmath}:j]} && \forall j \in \mathbb{Z}_E, \widetilde{\jmath} \in \mathbb{A}^{[j]}, \boldsymbol{x} \in \mathbb{W} : \|\boldsymbol{x}\|_2 = \rho \\ \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2 &= \phi_\eta^{[\widetilde{\jmath}:j]} && \forall j \in \mathbb{Z}_E, \widetilde{\jmath} \in \mathbb{A}^{[j]}, \boldsymbol{x} \in \mathbb{W} : \|\boldsymbol{x}\|_2 = r \\ \left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2 &= \phi_{\eta\downarrow}^{[j]} && \forall j \in \mathbb{Z}_E \cup \{0\}, \boldsymbol{x} \in \mathbb{X} : \|\boldsymbol{x}\|_2 = \rho \\ \left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2 &= \phi_\eta^{[j]} && \forall j \in \mathbb{Z}_E \cup \{0\}, \boldsymbol{x} \in \mathbb{X} : \|\boldsymbol{x}\|_2 = r \end{aligned} \tag{25}$$

C.1 Proof of Theorem 1 - Bilinear Representation

Recalling that the network is arranged in layers $\boldsymbol{\jmath} = 0, 1, 2, \ldots, D$, and given that we know the feature map representation for the input layer $\boldsymbol{\jmath} = 0$ is, tivially:

$$\boldsymbol{x}^{[0]} = \left\langle \boldsymbol{\Psi}^{[0]}(\Theta), \boldsymbol{\phi}^{[0]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[0]}}$$

where $\boldsymbol{\Psi}^{[0]}(\Theta) = r\mathbf{I}$, $\boldsymbol{\phi}^{[0]}(\boldsymbol{x}) = \boldsymbol{x}$ and $\mathbf{g}^{[0]} = \frac{1}{r}\mathbf{1}$, it suffices to show that *if* all outputs of all nodes $\widetilde{\jmath} \in \mathbb{L}^{[\boldsymbol{\jmath}-1]}$ ($\mathbb{L}^{[0]} = \{0\}$) in layer $\boldsymbol{\jmath} - 1$ can be expressed in terms of bilinear products:

$$\boldsymbol{x}^{[\widetilde{\jmath}]} = \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}]}} \tag{26}$$

then all nodes $j \in \mathbb{L}^{[\boldsymbol{\jmath}]}$, using the definitions given, can be written:

$$\boldsymbol{x}^{[\widetilde{\jmath}:j]} = \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}} \quad \forall \widetilde{\jmath} \in \mathbb{A}^{[j]} \tag{27}$$

and:

$$\boldsymbol{x}^{[j]} = \left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[j]}} \tag{28}$$

We call (27) the edge case and (28) the node case, and will treat them separately.

Edge case: We are given that (26) is correct. Substituting (21a) into the bilinear product and using (26), (3) and (17), we find that:

$$\begin{aligned} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}} &= \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}} \\ &= \boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0}) + \sum_{\mathtt{k} \succ_{H^{[\widetilde{\jmath}]}} 0} \mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]} \sum_{0 \prec_{H^{[\widetilde{\jmath}]}} \mathtt{l} \leq \mathtt{k}} \binom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \left[\left\langle \bigotimes_{\widetilde{\imath} \in \mathbb{N}_{H^{[\widetilde{\jmath}]}}} \boldsymbol{\Psi}_{:i_{\widetilde{\jmath}}}^{[\widetilde{\jmath}]}(\Theta)^{\otimes l_{i_{\widetilde{\jmath}}}}, \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})^{\otimes |\mathtt{l}|} \right]_{\mathbf{g}^{[\widetilde{\jmath}] \otimes |\mathtt{l}|}} \right]_{i_{\widetilde{\jmath}}} \\ &= \boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0}) + \sum_{\mathtt{k} \succ_{H^{[\widetilde{\jmath}]}} 0} \mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]} \sum_{0 \prec_{H^{[\widetilde{\jmath}]}} \mathtt{l} \leq \mathtt{k}} \binom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x}) \right]_{\mathbf{g}^{[\widetilde{\jmath}]}}^{\mathtt{l}} \\ &= \boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0}) + \sum_{\mathtt{k} \succ_{H^{[\widetilde{\jmath}]}} 0} \mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]} \sum_{0 \prec_{H^{[\widetilde{\jmath}]}} \mathtt{l} \leq \mathtt{k}} \binom{\mathtt{k}}{\mathtt{l}} \mathrm{He}_{\mathtt{k}-\mathtt{l}} \boldsymbol{x}^{[\widetilde{\jmath}]\mathtt{l}} \\ &= \boldsymbol{x}^{[\widetilde{\jmath}:j]} \end{aligned}$$

which is the desired result (27).

Node case: We have shown that (27) is correct. Substituting (21b) into the bilinear product and using (27), we find that, for columnar concatenation nodes $\bigcirc^{[j]} = \bigoplus$ (so $\square^{[j]} = \operatorname{diag}$, $\boxdot^{[j]} = \bigoplus$):

$$
\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right\rangle_{\mathbf{g}^{[j]}} &= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\square^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]}, \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \omega^{[\widetilde{\jmath}:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \begin{bmatrix} \frac{1}{\omega^{[\widetilde{\jmath}_1:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}_1:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_1:j]} & \mathbf{0} & \dots \\ \mathbf{0} & \frac{1}{\omega^{[\widetilde{\jmath}_2:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}_2:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_2:j]} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}, \begin{bmatrix} \omega^{[\widetilde{\jmath}_1:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}_1:j]}\left(\boldsymbol{x}\right) \\ \omega^{[\widetilde{\jmath}_2:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}_1:j]}\left(\boldsymbol{x}\right) \\ \vdots \end{bmatrix} \right\rangle \begin{bmatrix} \frac{1}{\omega^{[\widetilde{\jmath}_1:j]}} \mathbf{g}^{[\widetilde{\jmath}_1:j]} \\ \frac{1}{\omega^{[\widetilde{\jmath}_2:j]}} \mathbf{g}^{[\widetilde{\jmath}_2:j]} \\ \vdots \end{bmatrix} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \begin{bmatrix} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}_1:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_1:j]}, \boldsymbol{\phi}^{[\widetilde{\jmath}_1:j]}\left(\boldsymbol{x}\right) \right\rangle_{\mathbf{g}^{[\widetilde{\jmath}_1:j]}} \\ \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}_2:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_2:j]}, \boldsymbol{\phi}^{[\widetilde{\jmath}_2:j]}\left(\boldsymbol{x}\right) \right\rangle_{\mathbf{g}^{[\widetilde{\jmath}_2:j]}} \\ \vdots \end{bmatrix} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \begin{bmatrix} \mathbf{W}^{[\widetilde{\jmath}_1:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}_1:j]} \\ \mathbf{W}^{[\widetilde{\jmath}_2:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}_2:j]} \\ \vdots \end{bmatrix} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \underset{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}{\bigoplus} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
\end{aligned}
$$

For additive nodes $\bigcirc^{[j]} = \sum$ (so $\square^{[j]} = \bigoplus$, $\boxdot^{[j]} = \bigoplus$):

$$
\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right\rangle_{\mathbf{g}^{[j]}} &= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\square^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]}, \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \omega^{[\widetilde{\jmath}:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \begin{bmatrix} \frac{1}{\omega^{[\widetilde{\jmath}_1:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}_1:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_1:j]} \\ \frac{1}{\omega^{[\widetilde{\jmath}_2:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}_2:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}_2:j]} \\ \vdots \end{bmatrix}, \begin{bmatrix} \omega^{[\widetilde{\jmath}_1:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}_1:j]}\left(\boldsymbol{x}\right) \\ \omega^{[\widetilde{\jmath}_2:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}_1:j]}\left(\boldsymbol{x}\right) \\ \vdots \end{bmatrix} \right\rangle \begin{bmatrix} \frac{1}{\omega^{[\widetilde{\jmath}_1:j]}} \mathbf{g}^{[\widetilde{\jmath}_1:j]} \\ \frac{1}{\omega^{[\widetilde{\jmath}_2:j]}} \mathbf{g}^{[\widetilde{\jmath}_2:j]} \\ \vdots \end{bmatrix} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \sum_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]}, \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \sum_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
\end{aligned}
$$

For Kronecker-product nodes $\bigcirc^{[j]} = \bigotimes$ (so $\square^{[j]} = \bigotimes$, $\boxdot^{[j]} = \bigotimes$):

$$
\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right\rangle_{\mathbf{g}^{[j]}} &= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\square^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]}, \underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \omega^{[\widetilde{\jmath}:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\underset{\widetilde{\jmath}\in\mathbb{A}^{[j]}}{\boxdot^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]}, \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]} \right)^{\mathrm{T}} \left(\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right) \odot \left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{g}^{[\widetilde{\jmath}:j]} \right) \right) \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left(\tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right) \mathbf{W}^{[\widetilde{\jmath}:j]} \right)^{\mathrm{T}} \left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \odot \mathbf{g}^{[\widetilde{\jmath}:j]} \right) \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right) \right\rangle_{\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \bigotimes_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
\end{aligned}
$$

For Hadamard product nodes $\bigcirc^{[j]} = \bigodot$ (so $\square^{[j]} = \bigotimes^{\updownarrow}$, $\boxdot^{[j]} = \bigotimes$):

$$
\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}^{[j]}} &= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \square^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}, \boxdot^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]}\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\boxdot^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}}\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \bigotimes^{\updownarrow}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}, \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]}\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}}\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left(\bigotimes^{\updownarrow}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}\right)^{\mathrm{T}}\left(\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right)\odot\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right) \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left[\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}_{:i_j}\right)\right)^{\mathrm{T}}\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}_{:i_j}\right)^{\mathrm{T}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)^{\mathrm{T}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\left\langle\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right),\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \bigodot_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
\end{aligned}
$$

For multi-inner-product nodes $\bigcirc^{[j]} = \langle\!\langle\cdot\rangle\!\rangle$ (so $\square^{[j]} = \bigotimes^{\updownarrow}(\cdot)\,\mathbf{1}$, $\boxdot^{[j]} = \bigotimes$):

$$
\begin{aligned}
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}^{[j]}} &= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \square^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}, \boxdot^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]}\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\boxdot^{[j]}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}}\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left\langle \bigotimes^{\updownarrow}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}\mathbf{1}, \bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]}\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \frac{1}{\omega^{[\widetilde{\jmath}:j]}}\mathbf{g}^{[\widetilde{\jmath}:j]}} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \left(\bigotimes^{\updownarrow}_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \tfrac{1}{\omega^{[\widetilde{\jmath}:j]}} \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}\mathbf{1}\right)^{\mathrm{T}}\left(\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right)\odot\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right) \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\left[\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}_{:i_j}\right)\right)^{\mathrm{T}}\left(\bigotimes_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left(\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\mathbf{W}^{[\widetilde{\jmath}:j]}_{:i_j}\right)^{\mathrm{T}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)^{\mathrm{T}}\left(\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\odot\mathbf{g}^{[\widetilde{\jmath}:j]}\right)\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\left\langle\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right),\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\left[\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}_{:i_j}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right]_{i_j} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \mathbf{1}^{\mathrm{T}}\bigodot_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]} \\
&= \gamma^{[j]}\mathbf{b}^{[j]} + \bigl\langle\!\bigl\langle\cdot\bigr\rangle\!\bigr\rangle_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
\end{aligned}
$$

where on the final line we use $\langle\!\langle\cdot\rangle\!\rangle$ as an operator (see notation section). So, in all cases:

$$
\left\langle \boldsymbol{\Psi}^{[j]}\left(\Theta\right), \boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right]_{\mathbf{g}^{[j]}} = \gamma\mathbf{b}^{[j]} + \bigcirc^{[j]}_{\widetilde{\jmath}\in\widetilde{\mathbb{P}}^{[j]}}\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]} = \boldsymbol{x}^{[j]}
$$

which is the desired result (28) for the node case.

C.2 Proof of Theorem 1 - Norm-Bounds

Recalling that the network is arranged in layers $\boldsymbol{\jmath} = 0, 1, 2, \ldots, D$, and noting that for the input layer $\boldsymbol{\jmath} = 0$ is, tivially from our assumptions:

$$\begin{aligned} \left\|\boldsymbol{\Psi}^{[0]}\left(\Theta\right)\right\|_F &= \psi_\eta^{[0]} \\ \phi_{\downarrow\eta}^{[0]} \leq \left\|\boldsymbol{\phi}^{[0]}\left(\boldsymbol{x}\right)\right\|_F &= \phi_\eta^{[0]} \end{aligned}$$

where $\psi_\eta^{[0]} = r$, $\phi_{\downarrow\eta}^{[0]} = \rho$ and $\phi_\eta^{[0]} = r$, it suffices to show that *if* all outputs of all nodes $\widetilde{\jmath} \in \mathbb{L}^{[\boldsymbol{\jmath}-1]}$ in layer $\boldsymbol{\jmath} - 1$ satisfy:

$$\begin{aligned} \left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_F &= \psi_\eta^{[\widetilde{\jmath}]} \\ \phi_{\downarrow\eta}^{[\widetilde{\jmath}]} \leq \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}\left(\boldsymbol{x}\right)\right\|_F &= \phi_\eta^{[\widetilde{\jmath}]} \end{aligned} \tag{29}$$

then all nodes $j \in \mathbb{L}^{[\boldsymbol{\jmath}]}$, using the definitions given, satisfy:

$$\begin{aligned} \left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}\left(\Theta\right)\right\|_F &= \psi_\eta^{[\widetilde{\jmath}:j]} \\ \phi_{\downarrow\eta}^{[\widetilde{\jmath}:j]} \leq \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}\right)\right\|_F &= \phi_\eta^{[\widetilde{\jmath}:j]} \end{aligned} \qquad \forall \widetilde{\jmath} \in \mathbb{A}^{[j]} \tag{30}$$

and:

$$\begin{aligned} \left\|\boldsymbol{\Psi}^{[j]}\left(\Theta\right)\right\|_F &= \psi_\eta^{[j]} \\ \phi_{\downarrow\eta}^{[j]} \leq \left\|\boldsymbol{\phi}^{[j]}\left(\boldsymbol{x}\right)\right\|_F &= \phi_\eta^{[j]} \end{aligned} \tag{31}$$

We call (30) the edge case and (31) the node case, and will treat them separately.

Edge case: We are given that (29) is correct. By direct calculation, for incoming edges (21a), using the multinomial theorem at step $(*)$:

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 &= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \left\|\left[\operatorname{diag}\left(\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]\langle\odot\frac{1}{2}\rangle}\right)\otimes^{\updownarrow}\left[\binom{\mathtt{k}}{\mathbf{1}}^{\frac{1}{2}}\bigotimes_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\left(\sqrt{\eta}\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right)^{\otimes l_{\widetilde{\imath}}}\right]_{\substack{\mathbf{1}\succ_{H^{[\widetilde{\jmath}]}}0,\\ \mathbf{1}\leq\mathtt{k}}}\right]_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\right\|_F^2\right) \\
&= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left\|\left[\binom{\mathtt{k}}{\mathbf{1}}^{\frac{1}{2}}\bigotimes_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\left(\sqrt{\eta}\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right)^{\otimes l_{\widetilde{\imath}}}\right]_{\substack{\mathbf{1}\succ_{H^{[\widetilde{\jmath}]}}0,\\ \mathbf{1}\leq\mathtt{k}}}\right\|_2^2\right) \\
&= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\sum_{0\prec_{H^{[\widetilde{\jmath}]}}\mathbf{1}\leq\mathtt{k}}\binom{\mathtt{k}}{\mathbf{1}}\left\|\bigotimes_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\left(\sqrt{\eta}\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right)^{\otimes l_{\widetilde{\imath}}}\right\|_2^2\right) \\
&= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\sum_{0\prec_{H^{[\widetilde{\jmath}]}}\mathbf{1}\leq\mathtt{k}}\binom{\mathtt{k}}{\mathbf{1}}\prod_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\left(\sqrt{\eta}\left\|\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_2\right)^{2l_{\widetilde{\imath}}}\right) \\
&= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left(\left(\sum_{0\preceq_{H^{[\widetilde{\jmath}]}}\mathbf{1}\leq\mathtt{k}}\binom{\mathtt{k}}{\mathbf{1}}\mathbf{1}^{|\mathtt{k}|-|\mathbf{1}|}\prod_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\left(\eta\left\|\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_2^2\right)^{l_{\widetilde{\imath}}}\right) - 1\right)\right) \\
&=^{(*)} \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left(\left(1+\sum_{\widetilde{\imath}\in\mathbb{N}_{H^{[\widetilde{\jmath}]}}}\eta\left\|\boldsymbol{\Psi}_{:\widetilde{\imath}}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_2^2\right)^{|\mathtt{k}|} - 1\right)\right) \\
&= \frac{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\mathbf{0}\right)\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left(\left(1+\eta\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}\left(\Theta\right)\right\|_F^2\right)^{|\mathtt{k}|} - 1\right)\right) \\
&= L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}
\end{aligned}$$

and:

$$\begin{aligned}
\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2 &= \frac{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \left\|\left[\left|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right|^{\odot\frac{1}{2}} \otimes \left[\binom{\mathtt{k}}{1}^{\frac{1}{2}}\left(\sqrt{\eta}\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right)^{\otimes|\mathtt{1}|}\right]_{\mathtt{1}\succ_{H^{[\widetilde{\jmath}]}}0,\ \mathtt{1}\leq\mathtt{k}}\right]_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\right\|_2^2\right)\\
&= \frac{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left\|\left[\binom{\mathtt{k}}{1}^{\frac{1}{2}}\left(\sqrt{\eta}\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right)^{\otimes|\mathtt{1}|}\right]_{\mathtt{1}\succ_{H^{[\widetilde{\jmath}]}}0,\ \mathtt{1}\leq\mathtt{k}}\right\|_2^2\right)\\
&= \frac{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\sum_{0\prec_{H^{[\widetilde{\jmath}]}}\mathtt{1}\leq\mathtt{k}}\binom{\mathtt{k}}{1}\left\|\left(\sqrt{\eta}\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right)^{\otimes|\mathtt{1}|}\right\|_2^2\right)\\
&= \frac{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\sum_{0\prec_{H^{[\widetilde{\jmath}]}}\mathtt{1}\leq\mathtt{k}}\binom{\mathtt{k}}{1}\left(\sqrt{\eta}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2\right)^{2|\mathtt{1}|}\right)\\
&=^{(*)} \frac{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left(\eta^2\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}(\mathbf{0})\right\|_2^2 + \sum_{\mathtt{k}\succ_{H^{[\widetilde{\jmath}]}}0}\left\|\mathbf{a}_{\mathtt{k}}^{[\widetilde{\jmath}:j]}\right\|_1\left(\left(1+\eta\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)^{|\mathtt{k}|}-1\right)\right)\\
&= L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}
\end{aligned}$$

which we may bound as (using that $s_\eta^{[\widetilde{\jmath}:j]}$ is increasing on $\mathbb{R}^+$):

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 &= L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \leq \psi_\eta^{[\widetilde{\jmath}:j]2}\\
\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2 &= L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \in \left[\phi_{\downarrow\eta}^{[\widetilde{\jmath}:j]}, \phi_\eta^{[\widetilde{\jmath}:j]2}\right]
\end{aligned}$$

which is the desired result (30).

Node case: We have shown that (30) is correct. For columnar concatenation nodes $\bigcirc^{[j]} = \bigoplus$ (so $\square^{[j]} = \mathrm{diag}$, $\boxdot^{[j]} = \bigoplus$):

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 &= \left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\\
\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 &= \beta^{[j]2} + \sum_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\omega^{[\widetilde{\jmath}:j]2}
\end{aligned}$$

For additive nodes $\bigcirc^{[j]} = \sum$ (so $\square^{[j]} = \bigoplus$, $\boxdot^{[j]} = \bigoplus$):

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 &= \left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\\
\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 &= \beta^{[j]2} + \sum_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\omega^{[\widetilde{\jmath}:j]2}
\end{aligned}$$

For Kronecker-product nodes $\bigcirc^{[j]} = \bigotimes$ (so $\square^{[j]} = \bigotimes$, $\boxdot^{[j]} = \bigotimes$):

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 &= \left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\\
\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 &= \beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\omega^{[\widetilde{\jmath}:j]2}
\end{aligned}$$

For Hadamard product nodes $\bigcirc^{[j]} = \bigodot$ (so $\square^{[j]} = \bigotimes^{\updownarrow}$, $\boxdot^{[j]} = \bigotimes$):

$$\begin{aligned}
\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 &\leq \left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\\
\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 &= \beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)}\omega^{[\widetilde{\jmath}:j]2}
\end{aligned}$$

For multi-inner-product nodes $\bigcirc^{[j]} = \langle\!\langle \cdot \rangle\!\rangle$ (so $\square^{[j]} = \bigotimes^{\updownarrow}(\cdot)\,\mathbf{1}$, $\boxdot^{[j]} = \bigotimes$):

$$\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 \leq \left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2$$
$$\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 = \beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \omega^{[\widetilde{\jmath}:j]2}$$

Thus in general, for all nodes considered here:

$$\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 \leq \left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2$$
$$\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 = \beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} \omega^{[\widetilde{\jmath}:j]2}$$

which we may bound as:

$$\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2 \leq \psi_\eta^{[j]2}$$
$$\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2 \in \left[\phi_{\downarrow\eta}^{[j]2}, \phi_\eta^{[j]2}\right]$$

which is the desired result (31) for the node case.

We observe that the data-feature-map bound is tight:

$$\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2 = \phi_{\downarrow\eta}^2 \text{ if } \|\boldsymbol{x}\|_2 = \rho, \text{ and } \phi_{\downarrow\eta}^2 > 0 \text{ if } \rho > 0$$

$$\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2 = \phi_\eta^2 \text{ if } \|\boldsymbol{x}\|_2 = r$$

In the limit $\eta \to 0$, identifying $\psi^{[j]} = \psi_{0^+}^{[j]}$, $\phi_\downarrow^{[j]} = \phi_{\downarrow 0^+}^{[j]}$, $\phi^{[j]} = \phi_{0^+}^{[j]}$; $\psi = \psi_{0^+}$, $\phi_\downarrow = \phi_{\downarrow 0^+}$, $\phi = \phi_{0^+}$; $\psi = \psi^{[E]}$, $\phi_\downarrow = \phi_\downarrow^{[E]}$, $\phi = \phi^{[E]}$; where, recursively $\forall j \in \mathbb{Z}_E$:

$$\begin{aligned}
\psi^{[j]2} &\triangleq \beta^{[j]2} + \bigcirledcirc_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}^{[j]} \omega^{[\widetilde{\jmath}:j]2} \begin{cases} L_{\psi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\psi^{[\widetilde{\jmath}]2} & \text{if } \boldsymbol{\tau}^{[\widetilde{\jmath}:j]} \text{ Lipschitz} \\ B_{\psi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{\psi^{[\widetilde{\jmath}]2}}{\phi_\downarrow^{[\widetilde{\jmath}]2}} & \text{otherwise} \end{cases} \\
\phi_\downarrow^{[j]2} &\triangleq \beta^{[j]2} + \bigcirledcirc_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \begin{cases} L_{\psi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\phi_\downarrow^{[\widetilde{\jmath}]2} & \text{if } \boldsymbol{\tau}^{[\widetilde{\jmath}:j]} \text{ Lipschitz} \\ B_{\psi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} & \text{otherwise} \end{cases} \\
\phi^{[j]2} &\triangleq \beta^{[j]2} + \bigcirledcirc_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \begin{cases} L_{\phi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\phi^{[\widetilde{\jmath}]2} & \text{if } \boldsymbol{\tau}^{[\widetilde{\jmath}:j]} \text{ Lipschitz} \\ B_{\phi^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}\frac{\phi^{[\widetilde{\jmath}]2}}{\phi_\downarrow^{[\widetilde{\jmath}]2}} & \text{otherwise} \end{cases}
\end{aligned} \tag{32}$$

(here we have used that $\lim_{\eta\to 0} \frac{s_\eta^{[\widetilde{\jmath}:j]}(z)}{s_\eta^{[\widetilde{\jmath}:j]}(1)} = z$ by observation of the definition), which justifies our simplification in the main body of the paper.

C.3 Proof of Corollaries 2 and 3 - Continuity Bounds

Our approach here mimics the previous two proofs. For the input node $j = 0$, for a given $\Theta \in \mathbb{W}$:

$$\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[0]}(\Theta), \boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[0]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = C_{\Theta,\eta}^{[0]2} \triangleq 1 \quad \text{for given } \Theta \in \mathbb{W}$$
$$\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[0]}(\Theta), \boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[0]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = C_{\mathbb{W},\eta}^{[0]2} \triangleq 1 \quad \forall \Theta \in \mathbb{W}$$
$$\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[0]}(\Theta), \boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[0]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[0]}(\Theta)\right\|_F^2} = \frac{\|\boldsymbol{x}\|_2^2}{r^2} \leq C_{\boldsymbol{x},\eta}^{[0]2} \triangleq 1 \quad \text{for given } \boldsymbol{x} \in \mathbb{X}$$
$$\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[0]}(\Theta), \boldsymbol{\phi}^{[0]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[0]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[0]}(\Theta)\right\|_F^2} = \frac{\|\boldsymbol{x}\|_2^2}{r^2} \leq C_{\mathbb{X},\eta}^{[0]2} \triangleq 1 \quad \forall \boldsymbol{x} \in \mathbb{X}$$

As in the previous section consider a single node $j \in \mathbb{L}^{[\boldsymbol{j}]}$ in layer $\boldsymbol{j}$. Assume that, for all nodes in the previous layer $\widetilde{\jmath} \in \mathbb{L}^{[\boldsymbol{j}-1]}$:

$$\left.\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2} \leq C_{\Theta,\eta}^{[\widetilde{\jmath}]2} \quad \text{for given } \Theta \in \mathbb{W} \\
&\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2} \leq C_{\mathbb{W},\eta}^{[\widetilde{\jmath}]2} \quad \forall \Theta \in \mathbb{W} \\
&\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2} \leq C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}]2} \quad \text{for given } \boldsymbol{x} \in \mathbb{X} \\
&\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2} \leq C_{\mathbb{X},\eta}^{[\widetilde{\jmath}]2} \quad \forall \boldsymbol{x} \in \mathbb{X}
\end{aligned}\right\} \begin{aligned} &C_{\Theta,\eta}^{[\widetilde{\jmath}]2} \leq C_{\mathbb{W},\eta}^{[\widetilde{\jmath}]2} \\ &\\ &C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}]2} \leq C_{\mathbb{X},\eta}^{[\widetilde{\jmath}]2} \end{aligned}$$

Edge case: for a Lipschitz neural activation $\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}$, for incoming edges (21a), for fixed $\Theta \in \mathbb{W}$:

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}^{[\widetilde{\jmath}]}\right)\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}^{[\widetilde{\jmath}]}\right)\right\|_2^2}{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}](\boldsymbol{x})}\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{L_{C_{\Theta,\eta}^{[\widetilde{\jmath}]}\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{x}^{[\widetilde{\jmath}]}\right\|_2^2}{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}](\boldsymbol{x})}\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}} \\
&\leq \frac{L_{C_{\Theta,\eta}^{[\widetilde{\jmath}]}\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}} \left(\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2}{\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}](\boldsymbol{x})}\right\|_2^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}}\right) \left(\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2}\right) \\
&\leq \frac{L_{C_{\Theta,\eta}^{[\widetilde{\jmath}]}\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{L_{\phi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}} \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\|_2^2}
\end{aligned}$$

and similarly for fixed $\boldsymbol{x} \in \mathbb{X}$:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2} &= \sup_{\Theta\in\mathbb{W}} \frac{\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}^{[\widetilde{\jmath}]}\right)\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2} \\
&= \sup_{\Theta\in\mathbb{W}} \frac{\left\|\boldsymbol{\tau}^{[\widetilde{\jmath}:j]}\left(\boldsymbol{x}^{[\widetilde{\jmath}]}\right)\right\|_2^2}{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2} \frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}](\Theta)}\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}} \\
&\leq \sup_{\Theta\in\mathbb{W}} \frac{L_{C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}]}\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}} \frac{\left\|\boldsymbol{x}^{[\widetilde{\jmath}]}\right\|_2^2}{\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}](\Theta)}\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}} \\
&\leq \frac{L_{C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}]}\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}} \left(\sup_{\Theta\in\mathbb{W}} \frac{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2}{\frac{s_\eta^{[\widetilde{\jmath}:j]}\left(\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}](\Theta)}\right\|_F^2\right)}{s_\eta^{[\widetilde{\jmath}:j](1)}}}\right) \left(\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2}\right) \\
&\leq \frac{L_{C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}]}\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}}{L_{\psi_\eta^{[\widetilde{\jmath}]}}^{[\widetilde{\jmath}:j]2}} \sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta), \boldsymbol{\phi}^{[\widetilde{\jmath}]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{\jmath}]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}]}(\Theta)\right\|_F^2}
\end{aligned}$$

Alternatively, for bounded (non-Lipschitz) neural activations:

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{\tau}^{[\overline{j}:j]}\left(\boldsymbol{x}^{[\overline{j}]}\right)\right\|_2^2}{\frac{B^{[\overline{j}:j]2}}{\phi_{\downarrow\eta}^{[\overline{j}]2}} \frac{s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\phi}^{[\overline{j}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\overline{j}:j]}(1)}} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\frac{1}{B^{[\overline{j}:j]2}}\left\|\boldsymbol{\tau}^{[\overline{j}:j]}\left(\boldsymbol{x}^{[\overline{j}]}\right)\right\|_2^2}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} \frac{s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\phi}^{[\overline{j}]}(\boldsymbol{x})\right\|_2^2\right)}{s_\eta^{[\overline{j}:j]}(1)}} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{s_\eta^{[\overline{j}:j]}(1)}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\phi}^{[\overline{j}]}(\boldsymbol{x})\right\|_2^2\right)} \\
&=^{(*)} \frac{s_\eta^{[\overline{j}:j]}(1)}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} s_\eta^{[\overline{j}:j]}\left(\phi_{\downarrow\eta}^{[\overline{j}]2}\right)}
\end{aligned}$$

which is finite; and:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta)\right\|_F^2} &= \sup_{\Theta\in\mathbb{W}} \frac{\left\|\boldsymbol{\tau}^{[\overline{j}:j]}\left(\boldsymbol{x}^{[\overline{j}]}\right)\right\|_2^2}{\frac{B^{[\overline{j}:j]2}}{\phi_{\downarrow\eta}^{[\overline{j}]2}} \frac{s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\Psi}^{[\overline{j}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\overline{j}:j]}(1)}} \\
&= \sup_{\Theta\in\mathbb{W}} \frac{\frac{1}{B^{[\overline{j}:j]2}}\left\|\boldsymbol{\tau}^{[\overline{j}:j]}\left(\boldsymbol{x}^{[\overline{j}]}\right)\right\|_2^2}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} \frac{s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\Psi}^{[\overline{j}]}(\Theta)\right\|_F^2\right)}{s_\eta^{[\overline{j}:j]}(1)}} \\
&\leq \sup_{\Theta\in\mathbb{W}} \frac{s_\eta^{[\overline{j}:j]}(1)}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} s_\eta^{[\overline{j}:j]}\left(\left\|\boldsymbol{\Psi}^{[\overline{j}]}(\Theta)\right\|_F^2\right)}
\end{aligned}$$

which is unbounded in general. It follows that:

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\|_2^2} \leq C_{\Theta,\eta}^{[\overline{j}:j]2} \triangleq \begin{cases} \frac{L^{[\overline{j}:j]2}}{L^{[\overline{j}:j]2}} \frac{C_{\Theta,\eta}^{[\overline{j}]}\phi_\eta^{[\overline{j}]}}{\phi_\eta^{[\overline{j}]}} C_{\Theta,\eta}^{[\overline{j}]2} & \text{if } \boldsymbol{\tau}^{[\overline{j}:j]} \text{ is Lipschitz} \\ \frac{s_\eta^{[\overline{j}:j]}(1)}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} s_\eta^{[\overline{j}:j]}\left(\phi_{\downarrow\eta}^{[\overline{j}]2}\right)} & \text{otherwise} \end{cases} & \text{for given } \Theta\in\mathbb{W} \\
&\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\|_2^2} \leq C_{\mathbb{W},\eta}^{[\overline{j}:j]2} \triangleq \begin{cases} \frac{L^{[\overline{j}:j]2}}{L^{[\overline{j}:j]2}} \frac{C_{\mathbb{W},\eta}^{[\overline{j}]}\phi_\eta^{[\overline{j}]}}{\phi_\eta^{[\overline{j}]}} C_{\mathbb{W},\eta}^{[\overline{j}]2} & \text{if } \boldsymbol{\tau}^{[\overline{j}:j]} \text{ is Lipschitz} \\ \frac{s_\eta^{[\overline{j}:j]}(1)}{\frac{1}{\phi_{\downarrow\eta}^{[\overline{j}]2}} s_\eta^{[\overline{j}:j]}\left(\phi_{\downarrow\eta}^{[\overline{j}]2}\right)} & \text{otherwise} \end{cases} & \forall\Theta\in\mathbb{W} \end{aligned} \right\} C_{\Theta,\eta}^{[\overline{j}:j]} \leq C_{\mathbb{W},\eta}^{[\overline{j}]}$$

$$\begin{aligned}
&\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta)\right\|_F^2} \leq C_{\boldsymbol{x},\eta}^{[\overline{j}:j]2} \triangleq \begin{cases} \frac{L^{[\overline{j}:j]2}}{L^{[\overline{j}:j]2}} \frac{C_{\boldsymbol{x},\eta}^{[\overline{j}]}\psi_\eta^{[\overline{j}]}}{\psi_\eta^{[\overline{j}]}} C_{\boldsymbol{x},\eta}^{[\overline{j}]2} & \text{if } \boldsymbol{\tau}^{[\overline{j}:j]} \text{ is Lipschitz} \\ \infty & \text{otherwise} \end{cases} & \text{for given } \boldsymbol{x}\in\mathbb{X} \\
&\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta),\boldsymbol{\phi}^{[\overline{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\overline{j}:j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[\overline{j}:j]}(\Theta)\right\|_F^2} \leq C_{\mathbb{X},\eta}^{[\overline{j}:j]2} \triangleq \begin{cases} \frac{L^{[\overline{j}:j]2}}{L^{[\overline{j}:j]2}} \frac{C_{\mathbb{X},\eta}^{[\overline{j}]}\psi_\eta^{[\overline{j}]}}{\psi_\eta^{[\overline{j}]}} C_{\mathbb{X},\eta}^{[\overline{j}]2} & \text{if } \boldsymbol{\tau}^{[\overline{j}:j]} \text{ is Lipschitz} \\ \infty & \text{otherwise} \end{cases} & \forall\boldsymbol{x}\in\mathbb{X} \end{aligned} \right\} C_{\boldsymbol{x},\eta}^{[\overline{j}:j]} \leq C_{\mathbb{X},\eta}^{[\overline{j}:j]}$$

Edge case: using (21b), for columnar concatenation nodes $\bigcirc^{[j]} = \bigoplus$ (so $\square^{[j]} = \mathrm{diag}$, $\boxdot^{[j]} = \bigoplus$):

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{x}^{[j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigoplus\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigoplus\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \left\langle \boldsymbol{\Psi}^{[\widetilde{j}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{j}:j]}}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2 C_{\Theta,\eta}^{[\widetilde{j}:j]2}}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\Theta,\eta}^{[\widetilde{j}:j]2}\right\}\right\} \\
&= \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\Theta,\eta}^{[\widetilde{j}:j]2}\right\}\right\}
\end{aligned}$$

and:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq \sup_{\Theta\in\mathbb{W}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\Psi}^{[\widetilde{j}:j]}(\Theta)\right\|_F^2 C_{\boldsymbol{x},\eta}^{[\widetilde{j}:j]2}}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\boldsymbol{x},\eta}^{[\widetilde{j}:j]2}\right\}\right\} \\
&\leq \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\boldsymbol{x},\eta}^{[\widetilde{j}:j]2}\right\}\right\}
\end{aligned}$$

For additive nodes $\bigcirc^{[j]} = \sum$ (so $\square^{[j]} = \bigoplus$, $\boxdot^{[j]} = \bigoplus$):

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta), \boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{x}^{[j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{j}:j]}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]\mathrm{T}} \left\langle \boldsymbol{\Psi}^{[\widetilde{j}:j]}(\Theta), \boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[\widetilde{j}:j]}}\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{j}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2 C_{\Theta,\eta}^{[\widetilde{j}:j]2}}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2}{\beta^{[j]2} + \sum\limits_{\widetilde{j}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{j}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{j}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\Theta,\eta}^{[\widetilde{j}:j]2}\right\}\right\} \\
&= \max\left\{\gamma^{[j]}, \max_{\widetilde{j}\in\mathbb{A}^{[j]}} \left\{C_{\Theta,\eta}^{[\widetilde{j}:j]2}\right\}\right\}
\end{aligned}$$

and:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq \sup_{\Theta\in\mathbb{W}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \sum\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \max_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\{C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}\right\} \\
&\leq \max\left\{\gamma^{[j]}, \max_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\{C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}\right\}
\end{aligned}$$

For Kronecker-product nodes $\bigcirc^{[j]} = \bigotimes$ (so $\square^{[j]} = \bigotimes$, $\boxdot^{[j]} = \bigotimes$):

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{x}^{[j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigotimes\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigotimes\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}} \left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta),\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right\|_2^2}{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2 C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}}{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\beta^{[j]2} + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2} \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\} \\
&= \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\}
\end{aligned}$$

and:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq \sup_{\Theta\in\mathbb{W}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2 \left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\} \\
&\leq \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}
\end{aligned}$$

For Hadamard product nodes $\bigcirc^{[j]} = \bigodot$ (so $\square^{[j]} = \bigotimes^{\updownarrow}$, $\boxdot^{[j]} = \bigotimes$), using that the norm of the

Hadamard product of unit vectors is ≤ 1:

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{x}^{[j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigodot_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \bigodot_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta),\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2 C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\} \\
&= \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\}
\end{aligned}$$

and:

$$\begin{aligned}
\sup_{\Theta\in\mathbb{W}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq \sup_{\Theta\in\mathbb{W}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\} \\
&\leq \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}
\end{aligned}$$

For multi-inner-product nodes $\bigcirc^{[j]} = \langle\langle\cdot\rangle\rangle$ (so $\square^{[j]} = \bigotimes^{\updownarrow}(\cdot)\,\mathbf{1}$, $\boxdot^{[j]} = \bigotimes$), using that the multi-inner-product of (2-norm) unit vectors is at most 1:

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\left\langle \boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\boldsymbol{x}^{[j]}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \left\langle\left\langle \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\rangle\right\rangle_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]} + \left\langle\left\langle \mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\rangle\right\rangle_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\boldsymbol{x}^{[\widetilde{\jmath}:j]}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&= \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]\mathrm{T}}\left\langle \boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta),\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right]_{\mathbf{g}^{[\widetilde{\jmath}:j]}}\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2 + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2 C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\beta^{[j]2} + \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} \omega^{[\widetilde{\jmath}:j]2}\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2} \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\} \\
&= \max\left\{\gamma^{[j]}, \prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}} C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\}
\end{aligned}$$

and:

$$\begin{aligned}\sup_{\Theta\in\mathbb{W}}\frac{\left\|\left\langle\boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq \sup_{\Theta\in\mathbb{W}}\frac{\left\|\mathbf{b}^{[j]}\right\|_2^2+\prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\Psi}^{[\widetilde{\jmath}:j]}(\Theta)\right\|_F^2 C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}}{\left\|\mathbf{b}^{[j]}\right\|_2^2+\prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}\\ &\leq \frac{\left\|\mathbf{b}^{[j]}\right\|_2^2+\prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}{\left\|\mathbf{b}^{[j]}\right\|_2^2+\prod\limits_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}\left\|\mathbf{W}^{[\widetilde{\jmath}:j]}\right\|_2^2\left\|\boldsymbol{\phi}^{[\widetilde{\jmath}:j]}(\boldsymbol{x})\right\|_2^2}\max\left\{\gamma^{[j]},\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}\\ &\leq \max\left\{\gamma^{[j]},\prod_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\}\end{aligned}$$

Thus in general, for all nodes considered here:

$$\left.\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{X}}\frac{\left\|\left\langle\boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &\leq C_{\Theta,\eta}^{[j]2}\triangleq\max\left\{\gamma^{[j]},\bigdiamond_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}^{[j]}C_{\Theta,\eta}^{[\widetilde{\jmath}:j]2}\right\} &&\text{for given }\Theta\in\mathbb{W}\\ \sup_{\boldsymbol{x}\in\mathbb{X}}\frac{\left\|\left\langle\boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\|_2^2} &\leq C_{\mathbb{W},\eta}^{[j]2}\triangleq\max\left\{\gamma^{[j]},\bigdiamond_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}^{[j]}C_{\mathbb{W},\eta}^{[\widetilde{\jmath}:j]2}\right\} &&\forall\Theta\in\mathbb{W}\\ \sup_{\Theta\in\mathbb{W}}\frac{\left\|\left\langle\boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq C_{\boldsymbol{x},\eta}^{[j]2}\triangleq\max\left\{\gamma^{[j]},\bigdiamond_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}^{[j]}C_{\boldsymbol{x},\eta}^{[\widetilde{\jmath}:j]2}\right\} &&\text{for given }\boldsymbol{x}\in\mathbb{X}\\ \sup_{\Theta\in\mathbb{W}}\frac{\left\|\left\langle\boldsymbol{\Psi}^{[j]}(\Theta),\boldsymbol{\phi}^{[j]}(\boldsymbol{x})\right\rangle_{\mathbf{g}^{[j]}}\right\|_2^2}{\left\|\boldsymbol{\Psi}^{[j]}(\Theta)\right\|_F^2} &\leq C_{\mathbb{X},\eta}^{[j]2}\triangleq\max\left\{\gamma^{[j]},\bigdiamond_{\widetilde{\jmath}\in\mathbb{A}^{[j]}}^{[j]}C_{\mathbb{X},\eta}^{[\widetilde{\jmath}:j]2}\right\} &&\forall\boldsymbol{x}\in\mathbb{X}\end{aligned}\right\}\begin{aligned}&C_{\Theta,\eta}^{[j]}\leq C_{\mathbb{W},\eta}^{[j]}\\ &\\ &\\ &C_{\boldsymbol{x},\eta}^{[j]}\leq C_{\mathbb{X},\eta}^{[j]}\end{aligned}$$

where we have defined:

$$\bigdiamond^{[j]}=\begin{cases}\max & \text{if }\bigcirc^{[j]}\in\{\bigoplus,\sum\}\\ \prod & \text{if }\bigcirc^{[j]}\in\{\bigotimes,\langle\!\langle\cdot\rangle\!\rangle\}\end{cases}$$

Consequently, defining $C_{\Theta,\eta}=C_{\Theta,\eta}^{[E]}$, $C_{\mathbb{W},\eta}=C_{\mathbb{W},\eta}^{[E]}$, $C_{\boldsymbol{x},\eta}=C_{\boldsymbol{x},\eta}^{[E]}$, and $C_{\mathbb{X},\eta}=C_{\mathbb{X},\eta}^{[E]}$:

$$\left.\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{X}}\frac{\left\|\langle\boldsymbol{\Psi}(\Theta),\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathbf{g}}\right\|_2^2}{\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2} &\leq C_{\Theta,\eta}^2 &&\text{for given }\Theta\in\mathbb{W}\\ \sup_{\boldsymbol{x}\in\mathbb{X}}\frac{\left\|\langle\boldsymbol{\Psi}(\Theta),\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathbf{g}}\right\|_2^2}{\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2} &\leq C_{\mathbb{W},\eta}^2 &&\forall\Theta\in\mathbb{W}\\ \sup_{\Theta\in\mathbb{W}}\frac{\left\|\langle\boldsymbol{\Psi}(\Theta),\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathbf{g}}\right\|_2^2}{\left\|\boldsymbol{\Psi}(\Theta)\right\|_F^2} &\leq C_{\boldsymbol{x},\eta}^2 &&\text{for given }\boldsymbol{x}\in\mathbb{X}\\ \sup_{\Theta\in\mathbb{W}}\frac{\left\|\langle\boldsymbol{\Psi}(\Theta),\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathbf{g}}\right\|_2^2}{\left\|\boldsymbol{\Psi}(\Theta)\right\|_F^2} &\leq C_{\mathbb{X},\eta}^2 &&\forall\boldsymbol{x}\in\mathbb{X}\end{aligned}\right\}\begin{aligned}&C_{\Theta,\eta}\leq C_{\mathbb{W},\eta}\\ &\\ &C_{\boldsymbol{x},\eta}\leq C_{\mathbb{X},\eta}\end{aligned}$$

where $C_{\mathbb{W},\eta}$ is finite in general and $C_{\mathbb{X},\eta}$ is finite if all neural activations are Lipschitz.

The limit case $\eta\to 0^+$ is of particular interest here. Defining $C_\Theta=\lim_{\eta\to0^+}C_{\Theta,\eta}$, $C_\Theta^{[j]}=\lim_{\eta\to0^+}C_{\Theta,\eta}^{[j]}$, $C_{\mathbb{W}}=\lim_{\eta\to0^+}C_{\mathbb{W},\eta}$, $C_{\mathbb{W}}^{[j]}=\lim_{\eta\to0^+}C_{\mathbb{W},\eta}^{[j]}$, we observe that, using the form of the base case and recursion:

$$\begin{aligned}&C_\Theta=C_\Theta^{[j]}=C_{\mathbb{W}}=C_{\mathbb{W}}^{[j]}=1 && \text{if all neural activations are Lipschitz or bounded}\\ &C_{\boldsymbol{x}}=C_{\boldsymbol{x}}^{[j]}=C_{\mathbb{X}}=C_{\mathbb{X}}^{[j]}=1 && \text{if all neural activations are Lipschitz}\end{aligned}\qquad\forall j\in\mathbb{Z}_E$$

This result, combined with Theorem 1, suffices to prove Corollaries 2 and 3.

C.4 Bounds for Data Drawn from a Distribution

A common variation of our assumption $\boldsymbol{x}\in\mathbb{X}_{\rho,r}$ - that is, the assumption that $\boldsymbol{x}$ is hard-limited in terms of its 2-norm - is that $\boldsymbol{x}\sim\mathcal{X}$ is drawn from some data distribution $\mathcal{X}$. With regard to our analysis, for arbitrary data distributions it is not possible to extend our analysis; however if it can be proven that $\boldsymbol{x}\in\mathbb{X}_{\rho,r}$ with-high-probability $\geq 1-\epsilon$ for suitable ρ, r then our results will follow whp $\geq 1-\epsilon$. To take a simple example, suppose we draw data from an n-dimensional normal distribution:

$$\boldsymbol{x}\sim\mathcal{X}=\mathcal{N}\left(\mathbf{0}_n,\sigma^2\mathbf{I}_n\right)$$

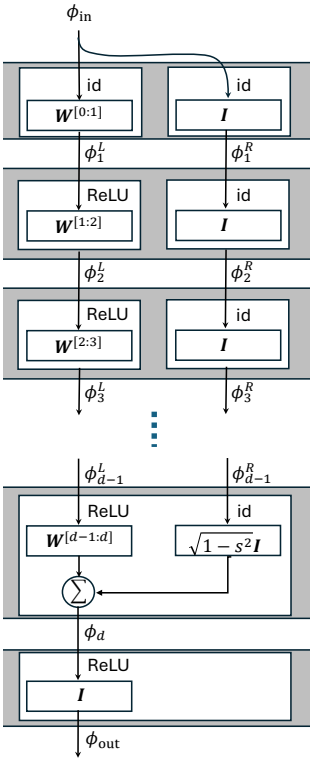


Figure 7: Calculation of ϕ_{out} in a residual network block.

Trivially, for $\boldsymbol{x} \sim \mathcal{X}$:

$$\Pr\left[\|\boldsymbol{x}\|_2 \leq \rho\right] \leq \frac{1}{2^{n/2}\sigma\Gamma\left(\frac{n}{2}+1\right)}\rho^n$$
$$\Pr\left[\|\boldsymbol{x}\|_2 \geq r\right] \leq 2e^{-\frac{r^2}{2n\sigma^2}}$$

Thus we have $\boldsymbol{x} \in \mathbb{X}_{\rho,r}$ with high probability $\geq 1-\epsilon$, where:

$$r = \sqrt{2n\ln\left(\frac{2}{(1-\upsilon)\epsilon}\right)}\sigma$$
$$\rho = \sqrt{2}\left(\Gamma\left(\frac{n}{2}+1\right)\upsilon\epsilon\right)^{\frac{1}{n}}\sigma$$

for some $\upsilon \in [0,1)$, In the purely Lipschitz case we can simplify this by setting $\upsilon = 0$ (so $\rho = 0$):

$$r = \sqrt{2n\ln\left(\frac{2}{\epsilon}\right)}\sigma \tag{33}$$

and more generally, if we allow non-Lipschitz neural activations, whp $\geq 1-\epsilon$:

$$\frac{r}{\rho} = \frac{\sqrt{n\ln\left(\frac{2}{(1-\upsilon)\epsilon}\right)}}{\left(\Gamma\left(\frac{n}{2}+1\right)\upsilon\epsilon\right)^{\frac{1}{n}}} \tag{34}$$

D Non-Trivial Blocks

In this section we consider norm- and continuity- bounds for particular common neural network archictectural blocks. Note that in all cases the continuity bounds C_Θ, $C_{\mathbb{W}}$, $C_{\boldsymbol{x}}$, $C_{\mathbb{X}}$ are well-behaved, so our task is to analyse the norm-bound ϕ. In this regard we refer the reader to (21) in Figure 6.

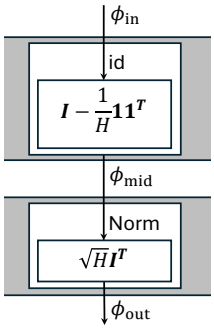


Figure 8: Calculation of ϕ_{out} in a residual network block.

D.1 Residual Block Bounds

In this section we consider the calculation of ϕ for a residual block. Figure 7 shows the notation we use here. All neural activations in this block are 1-Lipschitz so trivially, using our bounds:

$$\begin{aligned}
\psi^{\text{R}}_{d-1} &= \psi^{\text{R}}_{d-2} = \psi^{\text{R}}_{d-3} = \ldots = \psi^{\text{R}}_{2} = \psi^{\text{R}}_{1} = \psi_{\text{in}} \\
\phi^{\text{R}}_{d-1\downarrow} &= \phi^{\text{R}}_{d-2\downarrow} = \phi^{\text{R}}_{d-3\downarrow} = \ldots = \phi^{\text{R}}_{2\downarrow} = \phi^{\text{R}}_{1\downarrow} = \phi_{\text{in}\downarrow} \\
\phi^{\text{R}}_{d-1} &= \phi^{\text{R}}_{d-2} = \phi^{\text{R}}_{d-3} = \ldots = \phi^{\text{R}}_{2} = \phi^{\text{R}}_{1} = \phi_{\text{in}}
\end{aligned}$$

and:

$$\begin{aligned}
\psi^{\text{L}}_{d-1} &= \omega^{[d-2:d-1]}\psi^{\text{L}}_{d-2} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\psi^{\text{L}}_{d-3} = \ldots \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\psi^{\text{L}}_{1} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\omega^{[0:1]}\psi_{\text{in}} \\
\phi^{\text{L}}_{d-1\downarrow} &= \omega^{[d-2:d-1]}\phi^{\text{L}}_{d-2\downarrow} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\phi^{\text{L}}_{d-3\downarrow} = \ldots \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\phi^{\text{L}}_{1\downarrow} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\omega^{[0:1]}\phi_{\text{in}\downarrow} \\
\phi^{\text{L}}_{d-1} &= \omega^{[d-2:d-1]}\phi^{\text{L}}_{d-2} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\phi^{\text{L}}_{d-3} = \ldots \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\phi^{\text{L}}_{1} \\
&= \omega^{[d-2:d-1]}\omega^{[d-3:d-2]}\ldots\omega^{[1:2]}\omega^{[0:1]}\phi_{\text{in}}
\end{aligned}$$

and subsequently:

$$\begin{aligned}
\psi^{2}_{\text{out}} &= \psi^{2}_{d} \\
&= \psi^{\text{L}2}_{d-1} + \left(1-s^2\right)\psi^{\text{R}2}_{d-1} \\
&= \left(\omega^{[d-1:d]2}\ldots\omega^{[1:2]2}\omega^{[0:1]2} + 1 - s^2\right)\psi^{2}_{\text{in}} \\
\phi^{2}_{\text{out}\downarrow} &= \phi^{2}_{d\downarrow} \\
&= \phi^{\text{L}2}_{d-1\downarrow} + \left(1-s^2\right)\phi^{\text{R}2}_{d-1\downarrow} \\
&= \left(\omega^{[d-1:d]2}\ldots\omega^{[1:2]2}\omega^{[0:1]2} + 1 - s^2\right)\phi^{2}_{\text{in}\downarrow} \\
\phi^{2}_{\text{out}} &= \phi^{2}_{d} \\
&= \phi^{\text{L}2}_{d-1} + \left(1-s^2\right)\phi^{\text{R}2}_{d-1} \\
&= \left(\omega^{[d-1:d]2}\ldots\omega^{[1:2]2}\omega^{[0:1]2} + 1 - s^2\right)\phi^{2}_{\text{in}}
\end{aligned} \tag{35}$$

where we note that, for $\rho > 0$:

$$\frac{\phi_{\text{out}}}{\phi_{\text{out}\downarrow}} = \frac{\phi_{\text{in}}}{\phi_{\text{in}\downarrow}}$$

D.2 LayerNorm Block Bounds

As shown in Figure 8, the LayerNorm block is distinct insofar as it is non-Lipschitz. First we note that that $\|\sqrt{H}\mathbf{I}\|_2 = \sqrt{H}$, $\|\mathbf{I} - \frac{1}{H}\mathbf{1}\mathbf{1}^{\text{T}}\|_2 = 1$, so we may set $\omega^{[\text{in:mid}]} = 1$, $\omega^{[\text{mid:out}]} = \sqrt{H}$. Noting

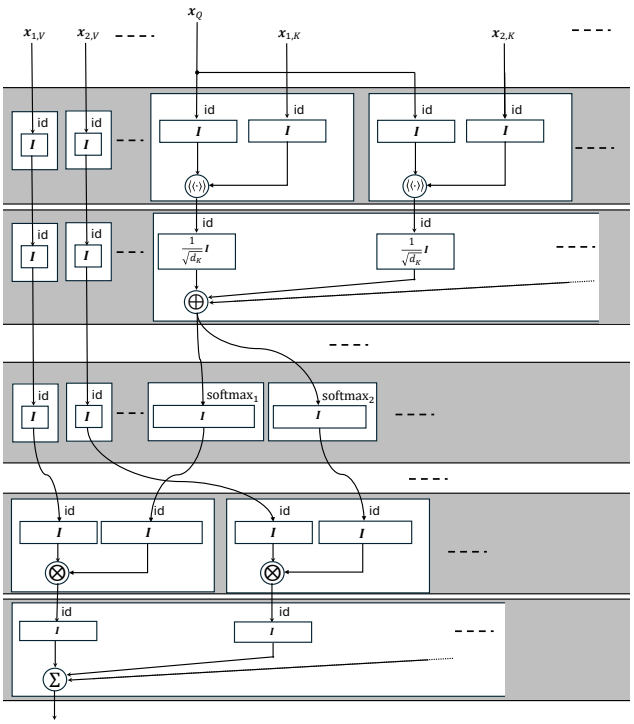


Figure 9: Single-query attention block.

that the Norm activation is non-Lipschitz and bounded by $B^{[\mathrm{Norm}]} = 1$, we see that:

$$\begin{aligned}\psi_{\mathrm{out}} &= \omega^{[\mathrm{mid:out}]} \tfrac{\psi_{\mathrm{mid}}}{\phi_{\downarrow \mathrm{mid}}} = \sqrt{H}\, \tfrac{\psi_{\mathrm{mid}}}{\phi_{\mathrm{mid}\downarrow}} \\ \phi_{\mathrm{out}\downarrow} &= \omega^{[\mathrm{mid:out}]} = \sqrt{H} \\ \phi_{\mathrm{out}} &= \omega^{[\mathrm{mid:out}]} \tfrac{\phi_{\mathrm{mid}}}{\phi_{\downarrow \mathrm{mid}}} = \sqrt{H}\, \tfrac{\phi_{\mathrm{mid}}}{\phi_{\mathrm{mid}\downarrow}}\end{aligned}$$

and trivially $\psi_{\mathrm{mid}} = \omega^{[\mathrm{in:mid}]}\psi_{\mathrm{in}} = \psi_{\mathrm{in}}$, $\phi_{\mathrm{mid}\downarrow} = \omega^{[\mathrm{in:mid}]}\phi_{\mathrm{in}\downarrow} = \phi_{\mathrm{in}\downarrow}$ and $\phi_{\mathrm{mid}} = \omega^{[\mathrm{in:mid}]}\phi_{\mathrm{in}} = \phi_{\mathrm{in}}$. Hence, overall:

$$\begin{aligned}\psi_{\mathrm{out}} &= \sqrt{H}\, \tfrac{\psi_{\mathrm{in}}}{\phi_{\mathrm{in}\downarrow}} \\ \phi_{\mathrm{out}\downarrow} &= \sqrt{H} \\ \phi_{\mathrm{out}} &= \sqrt{H}\, \tfrac{\phi_{\mathrm{in}}}{\phi_{\mathrm{in}\downarrow}}\end{aligned} \tag{36}$$

where we note that, for $\rho > 0$:

$$\tfrac{\phi_{\mathrm{out}}}{\phi_{\mathrm{out}\downarrow}} = \tfrac{\phi_{\mathrm{in}}}{\phi_{\mathrm{in}\downarrow}} \tag{37}$$

D.3 Single-query Attention Block Bounds

The standard bounds as presented in (21) are needlessly pessimistic for softmax nodes in attention blocks (Figure 2) as they are derived without taking into account the operation of the softmax in layer 3, which is a full softmax that has been split into components here - so while we can bound the set of *all* QK outputs, the standard bounds only bound the individual components without taking into account the interaction between then. The following more nuanced analysis gives a tighter bound.

In the following analysis we make the simplifying assumption $\psi_{\eta,\widetilde{\imath},V} = \psi_{\eta,V}$, $\phi_{\eta,\widetilde{\imath},V\downarrow} = \phi_{\eta,V\downarrow}$, $\phi_{\eta,\widetilde{\imath},V} = \phi_{\eta,V}$; $\psi_{\eta,\widetilde{\imath},K} = \psi_{\eta,K}$, $\phi_{\eta,\widetilde{\imath},K\downarrow} = \phi_{\eta,K\downarrow}$, $\phi_{\eta,\widetilde{\imath},K} = \phi_{\eta,K}$. Given this:

Layer 1: following the standard approach:

$$\begin{aligned}\psi^{[1]}_{\eta,\widetilde{\imath},V} &= \psi_{\eta,V} \\ \phi^{[1]}_{\eta,\widetilde{\imath},V\downarrow} &= \phi_{\eta,V\downarrow} \\ \phi^{[1]}_{\eta,\widetilde{\imath},V} &= \phi_{\eta,V}\end{aligned}$$

$$\begin{aligned}\psi^{[1]}_{\eta,\widetilde{\imath},QK} &= \psi_{\eta,Q}\psi_{\eta,K}\\ \phi^{[1]}_{\eta,\widetilde{\imath},QK\downarrow} &= \phi_{\eta,Q\downarrow}\phi_{\eta,K\downarrow}\\ \phi^{[1]}_{\eta,\widetilde{\imath},QK} &= \phi_{\eta,Q}\phi_{\eta,K}\end{aligned}$$

Layer 2: following the standard approach:

$$\begin{aligned}\psi^{[2]}_{\eta,\widetilde{\imath},V} &= \psi^{[1]}_{\eta,\widetilde{\imath},V} = \psi_{\eta,V}\\ \phi^{[2]}_{\eta,\widetilde{\imath},V\downarrow} &= \phi^{[1]}_{\eta,\widetilde{\imath},V\downarrow} = \phi_{\eta,V\downarrow}\\ \phi^{[2]}_{\eta,\widetilde{\imath},V} &= \phi^{[1]}_{\eta,\widetilde{\imath},V} = \phi_{\eta,V}\end{aligned}$$

$$\begin{aligned}\psi^{[2]2}_{\eta,QK} &= \tfrac{1}{d_K}\textstyle\sum_{\widetilde{\imath}}\psi^{[1]2}_{\eta,\widetilde{\imath},QK} = \psi^2_{\eta,Q}\psi^2_{\eta,K}\\ \phi^{[2]2}_{\eta,QK\downarrow} &= \tfrac{1}{d_K}\textstyle\sum_{\widetilde{\imath}}\phi^{[1]2}_{\eta,\widetilde{\imath},QK\downarrow} = \phi^2_{\eta,Q\downarrow}\phi^2_{\eta,K\downarrow}\\ \phi^{[2]2}_{\eta,QK} &= \tfrac{1}{d_K}\textstyle\sum_{\widetilde{\imath}}\phi^{[1]2}_{\eta,\widetilde{\imath},QK} = \phi^2_{\eta,Q}\phi^2_{\eta,K}\end{aligned}$$

Layer 3: we need to take some care with this layer. In particular, noting that the output of the layer is effectively the softmax split componentwise, we can constrain the sum of $\phi^{[3]}_{\eta,\widetilde{\imath},QK}$ as:

$$\begin{aligned}\psi^{[3]}_{\eta,\widetilde{\imath},V} &= \psi^{[2]}_{\eta,\widetilde{\imath},V} = \psi_{\eta,V}\\ \phi^{[3]}_{\eta,\widetilde{\imath},V\downarrow} &= \phi^{[2]}_{\eta,\widetilde{\imath},V\downarrow} = \phi_{\eta,V\downarrow}\\ \phi^{[3]}_{\eta,\widetilde{\imath},V} &= \phi^{[2]}_{\eta,\widetilde{\imath},V} = \phi_{\eta,V}\end{aligned}$$

$$\begin{aligned}\psi^{[3]2}_{\eta,\widetilde{\imath},QK} &= \lambda^2\tfrac{s^{[2:3]}_\eta\left(\psi^{[2]2}_{\eta,QK}\right)}{s^{[2:3]}_\eta(1)} = \lambda^2\tfrac{s^{[2:3]}_\eta\left(\psi^2_{\eta,Q}\psi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[3]2}_{\eta,\widetilde{\imath},QK\downarrow} &= c^2_{\widetilde{\imath}}\lambda^2\tfrac{s^{[2:3]}_\eta\left(\phi^{[2]2}_{\eta,QK\downarrow}\right)}{s^{[2:3]}_\eta(1)} = c^2_{\widetilde{\imath}}\lambda^2\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q\downarrow}\phi^2_{\eta,K\downarrow}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[3]2}_{\eta,\widetilde{\imath},QK} &= c^2_{\widetilde{\imath}}\lambda^2\tfrac{s^{[2:3]}_\eta\left(\phi^{[2]2}_{\eta,QK}\right)}{s^{[2:3]}_\eta(1)} = c^2_{\widetilde{\imath}}\lambda^2\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q}\phi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\end{aligned}$$

for some $c_1, c_2, \ldots \geq 0 : \sum_{\widetilde{\imath}} c^2_{\widetilde{\imath}} = 1$ (in the standard analysis we would let $c_1 = c_2 = \ldots = 1$).

Layer 4: following the standard approach:

$$\begin{aligned}\psi^{[4]2}_{\eta,\widetilde{\imath}} &= \psi^{[3]2}_{\eta,\widetilde{\imath},V}\psi^{[3]2}_{\eta,\widetilde{\imath},QK} = \lambda^2\psi^2_{\eta,V}\tfrac{s^{[2:3]}_\eta\left(\psi^2_{\eta,Q}\psi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[4]2}_{\eta,\widetilde{\imath}\downarrow} &= \phi^{[3]2}_{\eta,\widetilde{\imath},V\downarrow}\phi^{[3]2}_{\eta,\widetilde{\imath},QK\downarrow} = \lambda^2 c^2_{\widetilde{\imath}}\phi^2_{\eta,V\downarrow}\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q\downarrow}\phi^2_{\eta,K\downarrow}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[4]2}_{\eta,\widetilde{\imath}} &= \phi^{[3]2}_{\eta,\widetilde{\imath},V}\phi^{[3]2}_{\eta,\widetilde{\imath},QK} = \lambda^2 c^2_{\widetilde{\imath}}\phi^2_{\eta,V}\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q}\phi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\end{aligned}$$

Layer 5: recalling that $c_1, c_2, \ldots \geq 0$ satisfy $\sum_{\widetilde{\imath}} c^2_{\widetilde{\imath}} = 1$:

$$\begin{aligned}\psi^{[5]2}_{\eta} &= \textstyle\sum_{\widetilde{\imath}}\psi^{[4]2}_{\eta,\widetilde{\imath}} = d_K\lambda^2\psi^2_{\eta,V}\tfrac{s^{[2:3]}_\eta\left(\psi^2_{\eta,Q}\psi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[5]2}_{\eta,\downarrow} &= \textstyle\sum_{\widetilde{\imath}}\phi^{[4]2}_{\eta,\widetilde{\imath}\downarrow} = \lambda^2\phi^2_{\eta,V\downarrow}\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q\downarrow}\phi^2_{\eta,K\downarrow}\right)}{s^{[2:3]}_\eta(1)}\\ \phi^{[5]2}_{\eta} &= \textstyle\sum_{\widetilde{\imath}}\phi^{[4]2}_{\eta,\widetilde{\imath}} = \lambda^2\phi^2_{\eta,V}\tfrac{s^{[2:3]}_\eta\left(\phi^2_{\eta,Q}\phi^2_{\eta,K}\right)}{s^{[2:3]}_\eta(1)}\end{aligned}$$

Taking the limit $\eta \to 0^+$ we summarise the overall operation of this block as:

$$\begin{aligned}\psi_{\text{out}} &= \sqrt{d_K}\lambda\psi_{\text{in},V}\psi_{\text{in},Q}\psi_{\text{in},K}\\ \phi_{\text{out}\downarrow} &= \lambda\phi_{\text{in},V\downarrow}\phi_{\text{in},Q\downarrow}\phi_{\text{in},K\downarrow}\\ \phi_{\text{out}} &= \lambda\phi_{\text{in},V}\phi_{\text{in},Q}\phi_{\text{in},K}\end{aligned}$$

where we note that, for $\rho > 0$:

$$\tfrac{\phi_{\text{out}}}{\phi_{\text{out}\downarrow}} = \tfrac{\phi_{\text{in},V}}{\phi_{\text{in},V\downarrow}}\tfrac{\phi_{\text{in},Q}}{\phi_{\text{in},Q\downarrow}}\tfrac{\phi_{\text{in},K}}{\phi_{\text{in},K\downarrow}}$$

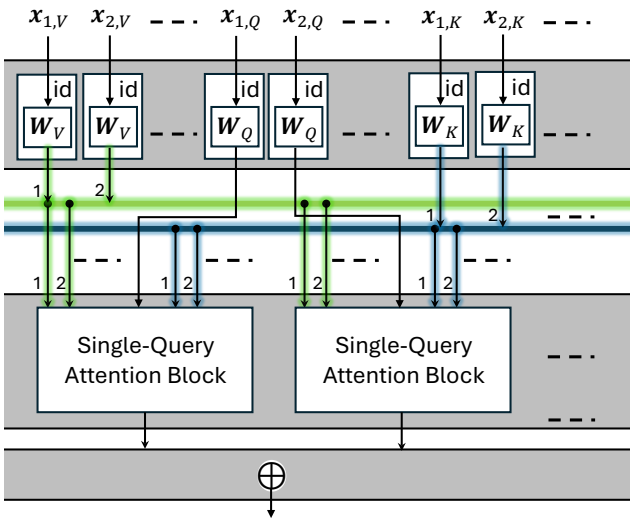


Figure 10: Single-Head attention block.

D.4 Single-Head and Multi-Head Attention Block Bounds

The standard single-head attention block is constructed from from single-query attention blocks as shown in figure 10. Multi-head attention is similar, with an additional h concatenations. Making the additional assumption, over section D.3, that $\psi_{\eta,\tilde{\imath},V} = \psi_{\eta,V}$, $\phi_{\eta,\tilde{\imath},V\downarrow} = \phi_{\eta,V\downarrow}$, $\phi_{\eta,\tilde{\imath},V} = \phi_{\eta,V}$, it is not difficult to see that:

$$\begin{aligned}\psi_{\text{out}} &= \sqrt{hd_K d_Q}\lambda\psi_{\text{in},V}\psi_{\text{in},Q}\psi_{\text{in},K}\\ \phi_{\text{out}\downarrow} &= \sqrt{hd_Q}\lambda\phi_{\text{in},V\downarrow}\phi_{\text{in},Q\downarrow}\phi_{\text{in},K\downarrow}\\ \phi_{\text{out}} &= \sqrt{hd_Q}\lambda\phi_{\text{in},V}\phi_{\text{in},Q}\phi_{\text{in},K}\end{aligned}$$

where d_Q is the number of queries and h is the number of heads. We note that, for $\rho > 0$:

$$\frac{\phi_{\text{out}}}{\phi_{\text{out}\downarrow}} = \frac{\phi_{\text{in},V}}{\phi_{\text{in},V\downarrow}}\frac{\phi_{\text{in},Q}}{\phi_{\text{in},Q\downarrow}}\frac{\phi_{\text{in},K}}{\phi_{\text{in},K\downarrow}} \tag{38}$$

E Bounds for Standard Network Toplogies

In this section we apply our results, and in particular our norm-bound $\|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq \phi\ \forall \boldsymbol{x} \in \mathbb{X}_{\rho,r}$ which is central in our Rademacher complexity bound, to standard network topologies.

E.1 Simple Unbiased Lipschitz Layerwise Network and ResNet

Consider a simple network with 1 unbiased node with L-Lipschitz activations per layer, so $D = E$, $j = \boldsymbol{\jmath} \in \mathbb{Z}_D$, and $\mathbb{A}^{[\boldsymbol{\jmath}]} = \{\boldsymbol{\jmath} - 1\}$. In this case, using (21), $\forall \boldsymbol{\jmath} \in \mathbb{Z}_D$:

$$\phi^{[\boldsymbol{\jmath}]} = L\omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]}\phi^{[\boldsymbol{\jmath}-1]}$$

and hence, using that $\phi^{[0]} = r$:

$$\phi = rL^D \textstyle\prod_{\boldsymbol{\jmath}\in\mathbb{Z}_D} \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]}$$

and we find that the norm-bound ϕ (and hence our Rademacher complexity bound) is proportional to the product of the weight-matrix norms, the maximum input norm r, and the exponentiated Lipschitz constant. In the distributional case, assuming $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ and using (33), whp $\geq 1 - \epsilon$:

$$\phi = \sigma L^D \sqrt{2n \ln\left(\tfrac{2}{\epsilon}\right)} \textstyle\prod_{\boldsymbol{\jmath}\in\mathbb{Z}_D} \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]}$$

We can also bound the residual network (ResNet) norm with this by including residual blocks as

nodes in the network. For example, if node $\boldsymbol{\jmath}$ is a residual block then the effective weight-norm bound $\omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]}$ becomes, for that non-trivial block, using (35):

$$\omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} = \left(\omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]_d 2} \dots \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]_2 2} \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]_1 2} + 1 - s^2\right)$$

where $\omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]_k}$ is the norm-bound for the k^{th} weight matrix $\mathbf{W}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]_k}$ in the residual block $\boldsymbol{\jmath}$.

E.2 Simple Unbiased non-Lipschitz Layerwise Network and LayerNorm

In this section we consider the same network as in the previous section E.1, excepting that we assume at least 1 non-Lipschitz, bounded neural activation. In this case, using (21), $\forall \boldsymbol{\jmath} \in \mathbb{Z}_D$:

$$\begin{aligned}
\phi_{\downarrow}^{[\boldsymbol{\jmath}]} &= \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \begin{cases} L_{\phi^{[\boldsymbol{\jmath}-1]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \phi_{\downarrow}^{[\boldsymbol{\jmath}-1]} & \text{if } \boldsymbol{\tau}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \text{ is Lipschitz} \\ B_{\phi^{[\bar{\jmath}]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} & \text{otherwise} \end{cases} \\
\phi^{[\boldsymbol{\jmath}]} &= \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \begin{cases} L_{\phi^{[\boldsymbol{\jmath}-1]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \phi^{[\boldsymbol{\jmath}-1]} & \text{if } \boldsymbol{\tau}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \text{ is Lipschitz} \\ B_{\phi^{[\boldsymbol{\jmath}-1]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \frac{\phi^{[\boldsymbol{\jmath}-1]}}{\phi_{\downarrow}^{[\boldsymbol{\jmath}-1]}} & \text{otherwise} \end{cases}
\end{aligned} \tag{39}$$

where $\phi_{\downarrow}^{[0]} = \rho$ and $\phi^{[0]} = r$. We immediately observe that:

$$\frac{\phi^{[D]}}{\phi_{\downarrow}^{[D]}} = \frac{\phi^{[D-1]}}{\phi_{\downarrow}^{[D-1]}} = \dots = \frac{\phi^{[0]}}{\phi_{\downarrow}^{[0]}} = \frac{r}{\rho} \qquad \forall \boldsymbol{\jmath} \in \mathbb{Z}_D$$

and hence (39) simplifies to:

$$\begin{aligned}
\phi_{\downarrow}^{[\boldsymbol{\jmath}]} &= \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \begin{cases} L_{\phi^{[\boldsymbol{\jmath}-1]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \phi_{\downarrow}^{[\boldsymbol{\jmath}-1]} & \text{if } \boldsymbol{\tau}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \text{ is Lipschitz} \\ B^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} & \text{otherwise} \end{cases} \\
\phi^{[\boldsymbol{\jmath}]} &= \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \begin{cases} L_{\phi^{[\boldsymbol{\jmath}-1]}}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \phi^{[\boldsymbol{\jmath}-1]} & \text{if } \boldsymbol{\tau}^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \text{ is Lipschitz} \\ B^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]} \frac{r}{\rho} & \text{otherwise} \end{cases}
\end{aligned} \tag{40}$$

If we further assume that node $\boldsymbol{\jmath} = D_{\downarrow}$ is the non-Lipschitz node closest to the output node, bounded by $B^{[D_{\downarrow}-1:D_{\downarrow}]}$, the norm bound becomes:

$$\phi = \frac{r}{\rho} B^{[D_{\downarrow}-1:D_{\downarrow}]} L^{D-D_{\downarrow}} \textstyle\prod_{\boldsymbol{\jmath}=D_{\downarrow}}^{D} \omega^{[\boldsymbol{\jmath}-1:\boldsymbol{\jmath}]}$$

The first thing to note with this bound is that it is no longer depth exponential, but rather depth-to-non-Lipschitz ($D - D_{\downarrow}$) exponential. This may appear surprising at first, but it is perhaps not so surprising when we note that the Lipschitz norm-bound scales with the max weight-matrix norm-bound, while a bounded neural-activation displays attributes that, in a crude sense, *flatten out* the magnitude of their input from previous layers. The obvious extreme case is a network combining ReLU and LayerNorm nodes, in which case we can scale weight matrices preceeding the LayerNorm arbitrarily without affecting the operation of the network in any way. This is directly reflected in the above expression, where the norm-bound ϕ is *independent* of the magnitude (matrix norm) of the weight-matrices in layers $1, 2, \dots, D_{\downarrow} - 1$ before the LayerNorm.

The ratio $\frac{r}{\rho}$ in the bound is perhaps less intuitive. In particular, while we would expect that the norm bound of $\|\boldsymbol{\phi}(\boldsymbol{x})\|_2$ should scale (increase) as $\|\boldsymbol{x}\|_2 \leq r$ increases (which the norm-bound does), it is less obvious that the bound should *increase* as the lower bound $\|\boldsymbol{x}\|_2 \geq \rho$ *decreases*. To understand this behaviour, recall that we only characterise neural activation $\boldsymbol{\tau}^{[D_{\downarrow}-1:D_{\downarrow}]}$ by its upper bound $B^{[D_{\downarrow}-1:D_{\downarrow}]}$ (1 for simplicity), so we must make a worst-case assumption that $\|\boldsymbol{x}^{[D_{\downarrow}]}\|_2 = 1$ for all $\boldsymbol{x} \in \mathbb{X}_{\rho,r}$. If $\|\boldsymbol{x}\|_2 = \rho$ then, in our worst-case analysis, the node must, in effect, *amplify* the input so that $\|\boldsymbol{x}^{[D_{\downarrow}]}\|_2 = 1$; the smaller ρ, the larger the amplification required.[9] This is why we take care not to over-claim in the case $\rho = r = 1$ in the main body of the paper.

Another apparent difficulty with this norm-bound is that one may argue that the lower bound $\|\boldsymbol{x}\|_2 \geq \rho$ is artificial, and that real data may not satisfy this bound. To cover this, we may use the distributional

[9] In the limit $\rho \to 0^+$ the amplication must approach ∞, which is why we insist $\rho > 0$ in this case.

case. For example, if node $D_\downarrow$ is a LayerNorm node and assuming $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I})$ then, using (36) and (34), with high probability $\geq 1-\epsilon$:

$$\phi = \frac{\sqrt{nH\ln\left(\frac{4}{\epsilon}\right)}}{\left(\Gamma\left(\frac{n}{2}+1\right)\frac{\epsilon}{2}\right)^{\frac{1}{n}}} L^{D-D_\downarrow} \prod_{\jmath=D_\downarrow}^{D} \omega^{[\jmath-1:\jmath]}$$

We observe that this bound is scale-independent, both in terms of the "size" σ of the data distribution and weight-norm bounds for layers prior to the final non-Lipschitz node $D_\downarrow$. The proportionality to $\sqrt{H}$ arises from the choice of LayerNorm, and the exact form of the new scaling arises from our choice of distribution.[10]

E.3 Transformers

Finally we may consider the Transformer. For concreteness we will assume the structure described in (Vaswani et al., 2017, Figure 1); and for tractability we will ignore the input/output embedding and positional encoding, and instead assume inputs and outputs (post-embedding/encoding) $\boldsymbol{x}_I, \boldsymbol{x}_O \in \mathbb{X}_{\rho,r} = \{\boldsymbol{x} \in \mathbb{R}^{d_K} : \rho \leq \|\boldsymbol{x}\|_2 \leq r\}$.

Encoder: The first layer in the encoder stack consists of a multi-head attention block inside a residual block, followed by a LayerNorm block. Using (38), the output norm-bound of the multihead attention block will satisfy:

$$\frac{\phi_{\text{mha}}}{\phi_{\text{mha}\downarrow}} = \left(\frac{\rho}{r}\right)^3$$

Subsequently, the output norm-bound of the residual block will satisfy:

$$\frac{\phi_{\text{res}}}{\phi_{\text{res}\downarrow}} = \frac{\phi_{\text{in}}+\phi_{\text{mha}}}{\phi_{\text{in}\downarrow}+\phi_{\text{mha}\downarrow}} = \frac{\phi_{\text{mha}}}{\phi_{\text{mha}\downarrow}} \frac{\frac{\phi_{\text{in}}}{\phi_{\text{mha}}}+1}{\frac{\phi_{\text{in}\downarrow}}{\phi_{\text{mha}\downarrow}}+1} = \left(\frac{\rho}{r}\right)^3$$

and we see from (37) that the output of the LayerNorm will satisfy:

$$\frac{\phi_{\text{m}}}{\phi_{\text{m}\downarrow}} = \left(\frac{\rho}{r}\right)^3$$

This is followed by a feed-forward network inside a residual block, again followed by a LayerNorm block. The analysis of this is similar to the above, excepting that, because the block inside the residual block is additive, there is no need to cube the ratio. The output of the LayerNorm in this layer will therefore satisfy:

$$\frac{\phi_{\text{l}}}{\phi_{\text{l}\downarrow}} = \left(\frac{\rho}{r}\right)^3$$

At total of[11] $M = 6$ of these layers occur sequentially, where for each the ratio is cubed due to the presence of the multi-head attention block. Subsequently, for the output of the encoder, we find:

$$\frac{\phi_{\text{enc}}}{\phi_{\text{enc}\downarrow}} = \left(\frac{\rho}{r}\right)^{3^M}$$

Decoder: The decoder is similar, with some important caveats. Perhaps most importantly, in the first layer the output of the second attention block (and therefore the output of the first layer in the decoder) will satisfy:

$$\frac{\phi_{\text{MHA}}}{\phi_{\text{MHA}\downarrow}} = \left(\frac{\rho}{r}\right)^{3+2.3^M} \leq \left(\frac{\rho}{r}\right)^{3^{M+1}}$$

This is followed by $(M-1) = 5$ additional layers, and so it may be seen that the output of the decoder, prior to the final linear and softmax, will satisfy:

$$\frac{\phi_{\text{dec}}}{\phi_{\text{dec}\downarrow}} \leq \left(\frac{\rho}{r}\right)^{3^{M+1}3^{2M-2}} = \left(\frac{\rho}{r}\right)^{3^{3M-1}}$$

and, using (36):

$$\phi_{\text{dec}} \leq \left(\frac{\rho}{r}\right)^{3^{3M-1}} \sqrt{d_{\text{model}}}$$

Subsequently, assuming the weights in the linear output layer of the Transformer satisfy $\|\mathbf{W}\|_2 \leq \omega$ and assuming $\lambda = 1$ in the final softmax we find that the overall norm-bound for the Transformer is:

$$\phi \leq \sqrt{d_{\text{model}}}\omega \left(\frac{\rho}{r}\right)^{3^{3M-1}}$$

[10] It may be informative to investigate the impact of the distribution $\boldsymbol{x} \sim \mathcal{X}$ on this bound in future work.

[11] We use M here rather than N due to the notational clash between (Vaswani et al., 2017) and our use of N.

F Proof of Theorem 4 - Rademacher Complexity

We are concerned with calculating the Rademacher complexity of:

$$\mathcal{R}_N\left(\mathcal{F}\right) = \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta\in\mathbb{W}} \tfrac{1}{N}\sum_{i,k} \epsilon_k h\left(\mathbf{f}\left(\boldsymbol{x}_k\right)\right)\right]$$

where h is L-Lipschitz. We have from (Maurer, 2016) that:

$$\mathcal{R}_N\left(\mathcal{F}\right) \leq \sum_i \sqrt{2} L \mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta\in\mathbb{W}} \tfrac{1}{N}\sum_k \epsilon_k f_i\left(\boldsymbol{x}_k\right)\right]$$

Thus we reduce the dimensionality of the problem to 1-dimension. Proceeding with the standard argument:

$$\begin{aligned}
\mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sup_{\Theta\in\mathbb{W}} \tfrac{1}{N}\sum_k \epsilon_k f_i\left(\boldsymbol{x}_k\right)\right] &= \tfrac{1}{N}\mathbb{E}_\nu \mathbb{E}_\epsilon \left[\sqrt{\sup_{\Theta\in\mathbb{W}} \left(\sum_k \epsilon_k f_i\left(\boldsymbol{x}_k\right)\right)^2}\right] \\
&\leq^{\text{Jensen's}-\text{inequality}} \tfrac{1}{N}\mathbb{E}_\nu \left[\sqrt{\mathbb{E}_\epsilon \sup_{\Theta\in\mathbb{W}} \left(\sum_k \epsilon_k f_i\left(\boldsymbol{x}_k\right)\right)^2}\right] \\
&= \tfrac{1}{N}\mathbb{E}_\nu \left[\sqrt{\mathbb{E}_\epsilon \sup_{\Theta\in\mathbb{W}} \left(\sum_{k,l} \epsilon_k \epsilon_l f_i\left(\boldsymbol{x}_k\right) f_i\left(\boldsymbol{x}_l\right)\right)}\right] \\
&=^{\mathbb{E}_\epsilon \epsilon_k \epsilon_l = \delta_{k,l}} \tfrac{1}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} f_i^2\left(\boldsymbol{x}_k\right)}\right] \\
&= \tfrac{1}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} \left\langle \boldsymbol{\Psi}_{:i}\left(\Theta\right), \boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right]_{\mathbf{g}}^2}\right] \\
&= \tfrac{1}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} \left(\tfrac{\left\langle \boldsymbol{\Psi}_{:i}\left(\Theta\right), \boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right]_{\mathbf{g}}}{\left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2}\right)^2 \left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2}\right]
\end{aligned}$$

and so:

$$\begin{aligned}
\mathcal{R}_N\left(\mathcal{F}\right) &\leq \tfrac{\sqrt{2}L}{N}\mathbb{E}_\nu \left[\sum_i \sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} \left(\tfrac{\left\langle \boldsymbol{\Psi}_{:i}\left(\Theta\right), \boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right]_{\mathbf{g}}}{\left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2}\right)^2 \left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2}\right] \\
&\leq^{1-\text{norm}-2-\text{norm}-\text{inequality}} \tfrac{\sqrt{2m}L}{N}\mathbb{E}_\nu \left[\sqrt{\sum_i \sum_k \sup_{\Theta\in\mathbb{W}} \left(\tfrac{\left\langle \boldsymbol{\Psi}_{:i}\left(\Theta\right), \boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right]_{\mathbf{g}}}{\left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2}\right)^2 \left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2}\right] \\
&= \tfrac{\sqrt{2m}L}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} \tfrac{\left\|\left\langle \boldsymbol{\Psi}_{:i}\left(\Theta\right), \boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right]_{\mathbf{g}}\right\|_2^2}{\left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2} \left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2}\right] \\
&=^{\text{bilinear}-\text{continuity}} \tfrac{\sqrt{2m}L}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k \sup_{\Theta\in\mathbb{W}} C_{\Theta,\eta}^2 \left\|\boldsymbol{\phi}\left(\boldsymbol{x}_k\right)\right\|_2^2}\right] \\
&=^{\text{norm}-\text{bounding}} \tfrac{\sqrt{2m}L}{N}\mathbb{E}_\nu \left[\sqrt{\sum_k C_{\mathbb{W},\eta}^2 \phi_\eta^2}\right] =^{\text{cleanup}} \tfrac{\sqrt{2m}L}{\sqrt{N}} C_{\mathbb{W},\eta}\phi_\eta
\end{aligned}$$

and the final result follows in the limit $\eta \to 0^+$, recalling $\lim_{\eta\to 0^+} C_{\mathbb{W},\eta} = 1$, $\lim_{\eta\to 0^+}\phi_\eta = \phi$:

$$\mathcal{R}_N\left(\mathcal{F}\right) \leq \tfrac{\sqrt{2m}L}{\sqrt{N}}\phi$$

NB: in the special case $m = 1$, $h = \mathrm{id}$, we can skip the first step which contributed the factor $\sqrt{2}L$ and the 1-norm-2-norm-inequality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the abstract/introduction were written after the key contributions were completed specifically to reflect them.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: the limitations of the work are clearly outlined in the Setting and Assumptions section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (eg., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, eg., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are clearly stated in the body of the paper. Most (non-trivial) proofs are summarised in the body, with reference to relevant appendices for details.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This is a purely theoretical work. Results will apply to any network satisfying our assumption, which are analytic in nature: network topology, bounds on weights/biases that are translated to the final result, and requirements on neural network activation functions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (eg., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (eg., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This is a purely theoretical work. As noted previously, results will apply to any network satisfying our assumption, which are analytic in nature: network topology, bounds on weights/biases that are translated to the final result, and requirements on neural network activation functions.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: See previous justification re results, data and code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: See previous justification re results, data and code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: See previous justification re results, data and code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is purely theory, so I cannot foresee specific societal impacts beyond improved performance in neural networks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: see previous responses.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: this is a theory paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: this is a theory paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are not core to this method, though the complexity bounds derived herein may apply to them.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.